A SIAMESE CONTENT-ATTENTIVE GRAPH CONVOLUTIONAL NETWORK FOR PERSONALITY RECOGNITION USING PHYSIOLOGY

Hao-Chun Yang, 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

Affective multimedia content has long been used as stimulation to study an individual's personality using physiology. In this work, we propose a novel Siamese Content-Attentive Graph Convolutional Network (SCA-GCN) to learn a discriminative physiology representation jointly guided by the actual video content of the emotional stimuli. The visual content of the stimuli is integrated into learning to weight the importance of physiology in the task of personality recognition. We evaluate our framework on a large public corpus of physiological data. Our method achieves the state of the art unweighted accuracy of 72.1%,69.5%, and 68.2% in a binary classification for dimensions of Openness, Emotion Stability, and Extraversion, which improves over the baseline DNN by 20.4%, 9%, and 13.9%. Further analysis reveals that there indeed exists a substantial effect from the media content in affecting the subject's internal physiological responses that result in an improved personality recognition performances.

Index Terms— affective multimedia, personality recognition, physiology, graph convolution network

1. INTRODUCTION

Personality has long been regarded as a key psychological construct that can be characterized into a few stable and measurable attributes due to its role in influencing an individual's emotion, modulating behaviors, and triggering decisions. Research has pointed out there exists a differential impact of affective media content on humans as a function of an individual's personality. For example, there is a significant preference bias for extroverts on the choice of TV programs and music genre [1]; individuals with higher Openness personality traits often favor reflective/complex music (such as jazz), while people with higher Neuroticism prefer more emotional music [2]. In [3] also demonstrates that visual patterns extracted from 'favorite' Flickr images can be used to predict user traits. This intriguing connection between personality and media data suggests that the content itself could act as an external indicator for uncovering a subject's personality traits. Developing algorithms to automatically infer a subject's personality, i.e., Automatically Personality Recognition (APR), has become crucial also in delivering personalized media content with impact [4].

Past research on APR has largely been developed in modeling different signal modalities. For example, many pieces of research have studied APR using lexical information [5] with many have mostly found its application on the social media platform to enable personalized profiling [6]. Recently, the proliferation of miniaturized sensors has enabled precise monitoring of various human internal physiological signals. In contrast to expressive cues (such as text), these biomarkers provide a scientifically grounded indicator to model personality traits directly from neurophysiological evidence.

Most if not all of these works share a common experimental setting, that is, by using emotion-rich audio-video media data as stimuli to elicit subject's internal physiological responses, then these physiological measurements are further processed for automatic recognition of personality traits. In fact, APR developments have almost all been operated in this particular setting [7, 8]. Recently, aside from modeling solely the physiology of the subject of interest, a couple of works have incorporated other meta information for joint APR modeling, e.g., fusing reactive expressions [9] or including the emotion variation [10]. However, these works neglect the fact that an individual's internal physiology is triggered directly by the displayed audio-visual content, which serves as a latent conditional control toward internal physiology. Hence, we argue that to develop an enhanced APR model, these media content signals should be integrated to properly model the intricate dependencies of personality as a function of affective media stimuli and physiological responses.

Specifically, we propose a novel Siamese Content-Attentive Graph Convolutional Network (SCA-GCN) for personality recognition using physiology. Our framework is evaluated on a publicly available large-scale AMIGOS dataset [8], in which each subject is exposed to a set of audio-visual movie clips with varying degree of intended affect-triggering content. We jointly model how a person's internal physiology response to these multiple stimuli with a graph structure, then further incorporate the media descriptors as attention modulation on the learned subject-wise graph-embedding of physiological signals. We achieve an unweighted recall of 72.1% on Openness, 69.5% on Emotion Stability, and 68.2% on Extraversion which is 20.4%, 9%, and 13.9% relative improvement over the vanilla DNN method.

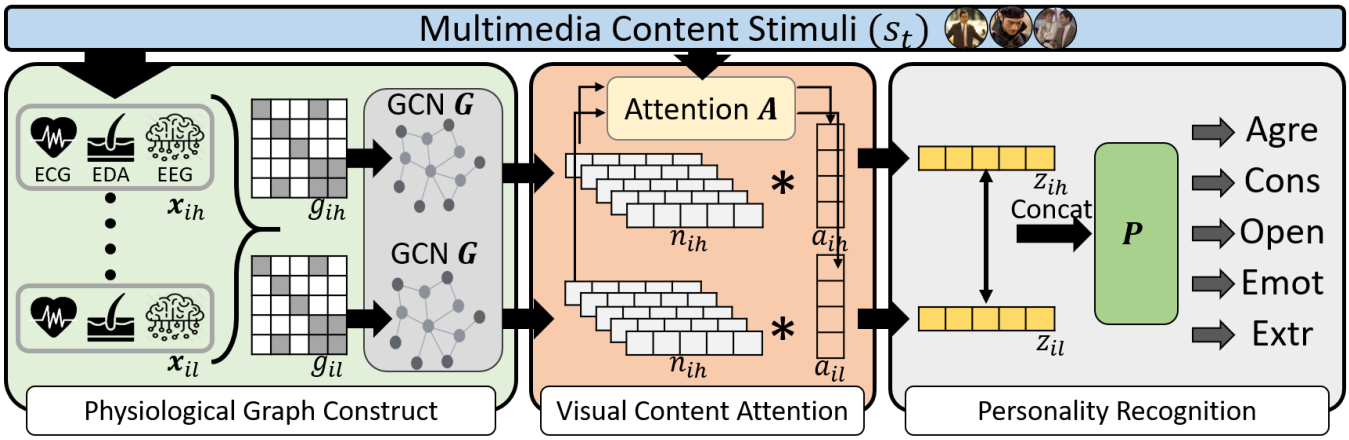

Fig. 1. Our proposed Siamese Content-Attentive Graph Convolution Network

2. RESEARCH METHODOLOGY

2.1. AMIGOS Dataset

We use the AMIGOS Dataset [8] for algorithm development and evaluation. A total of 16 short emotional videos (duration<250s) were carefully chosen from previous research as physiology elicitation. Each of the videos had intended *emotional* effect (indicated as *high/low* level for both arousal and valence dimensions) when being displayed to a subject. For each subject, the Big-Five personality[1] is measured using an online questionnaire, which maps personal traits into five dimensions [11]. During the experiment, each subject's ECG, EDA, and EEG signal are recorded simultaneously. Hence, there are three different physiological modalities recorded for each subject of every video stimuli, and our goal is to infer each subject's Big-Five personality traits using these physiological measurements collected during the 16 video clips. We carry out our personality recognition experiments as a binary classification problem (for each of the five personality traits) using cut-off at the median of all subjects.

2.2. Computational Framework

2.2.1. Physiological Descriptors and Graph Building

For preprocessing physiology data, a bandpass filter from 4-45Hz is applied on EEG while a low-pass filter cut-off at 60Hz is applied on ECG and EDA data. Several standard low-level physiological descriptors (LLDs) are listed in Table 1 and extracted using the NeuroKit [12]. A standard z-normalization is performed subject-wise on each feature dimension to mitigate the issue of individual differences.

We utilize a graph structure to encode the structural relationship of a subject's physiological responses (LLDs) across 16 different emotional video stimulus. Specifically, consider a set of subject i's d-dimensional LLDs $\boldsymbol{x_i} = \{x_i^1, \ldots, x_i^n\} \subset \mathbb{R}^d$ while n denotes the number of the stimuli during the experiment, subgraph $\mathcal{G}_{ic}$ are extracted from $\boldsymbol{x_{ic}} \subset \boldsymbol{x_i}$ where $c \subset \{h, l\}$ indicating whether the sample belong to *high* or *low* level of the original stimulation's intended emotional effect. The nodes of the graph $\mathcal{G}_{ic}$ are the extracted LLDs $\boldsymbol{x_{ic}}$, and the Pearson correlations are calculated from any of the two nodes and those larger than zero are connected as the edges of the graph. This procedure results in having two physiological graphs $\mathcal{G}_{ih}$ and $\mathcal{G}_{il}$ per subject.

Table 1. An overview of physiological low-level descriptors extracted from [12]. "F*" indicates 15 statistical functions[2]. EEG features are calculated for each channel then concatenated as a single feature vector.

Modality	Low-Level Descriptors
EEG(378)	Hjorth, Kurtosis, Skewness, First_diff_mean, First_diff_max, Sec_diff_mean, Sec_diff_max, Slope_mean, Slope_var, Wavelets, MaxPwelch, Entropy, ARMPB
ECG(51)	number_of_artifacts, RMSSD, meanNN, sdNN, cvNN, CVSD, medianNN, madNN, mcvNN, pNN50, pNN20, Triang, Shannon_h, ULF, VLF, LF, HF, VHF, Total_Power, LFn,HFn, LF/HF, LF/P, HF/P, DFA_1, DFA_2,Shannon, FD_Higushi, Average_Signal_Quality, F* Cardiac_Cycles_Signal_Quality
EDA(68)	F*SCR_Onsets, F*SCR_Peaks_Amplitudes, F*EDA_Phasic, F*EDA_Tonic_Component

2.2.2. Graph Convolutional Network (GCN)

Our model is primarily motivated as an extension of the Graph Convolutional Model (GCN) that performs a spectral convolution on first-order graph neighborhoods [13]. Recently GCN has received growing attention for its power on capturing inter-relationship between instances (nodes) [14]. The core GCN layer can be interpreted as a special case of a simple differentiable message-passing framework:

$$H^{(l+1)} = \sigma(\tilde{\mathrm{D}}^{-\frac{1}{2}}\tilde{\mathrm{A}}\tilde{\mathrm{D}}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \quad (1)$$

[1]Agre: Agreeableness, Cons: Conscientiousness, Open: Openness, Emot: Emotional Stability, Extr: Extraversion

[2]max, min, mean, median, std, skewness, kurtosis, min position, max position, 25_percentile, 75_percentile, 75_percentile-25_percentile, 1_percentile, 99_percentile, 99_percentile-1_percentile

Table 2. A summary of recognition results. '-c': concatenate a subject's all physiological responses as a single vector for classification [8]; '-v': predict personality though majority voting of each response; Aro: Arousal; Val: Valence. Both of these two indicate the emotional reference during sub-graph splitting. The chance UAR is 0.5.

	SVM-c	SVM-v	DNN-v	AMIL	G-1-a	G-1-b	G-1-c	G-1-d	G-2-a		G-2-b		G-2-c		G-2-d	
									Aro	Val	Aro	Val	Aro	Val	Aro	Val
Agre	0.500	0.389	0.534	0.563	0.500	0.527	0.510	0.532	0.558	0.566	0.563	0.605	0.574	**0.642***	0.558	0.603
Cons	0.455	0.406	0.549	0.508	0.507	0.509	0.489	0.524	0.553	0.537	0.563	**0.579***	0.537	0.566	0.526	0.526
Open	0.500	0.452	0.517	0.529	0.505	0.510	0.512	0.593	0.647	0.655	0.674	0.676	0.682	0.674	0.676	**0.721***
Emot	0.473	0.553	0.605	0.553	0.611	0.608	0.613	0.618	0.637	0.618	0.671	0.624	0.689	0.624	**0.695***	0.650
Extr	0.500	0.509	0.543	0.538	0.602	0.583	0.585	0.587	0.679	0.656	0.676	0.663	0.668	0.671	**0.682***	0.661

Here, H^l denotes the l^{th} layer in the network, σ is the non-linearity, W is the learnable weight matrix of shape $d^l \text{x} d^{l+1}$, and D,A refers to degree and adjacency matrix respectively. The $\sim$ is a renormalization trick in which that the self-connection is added to each node of the graph. The shape of the input H^0 is $N \times d$, where N is the number of nodes. This formula can be implemented and backpropagated using sparse matrix multiplication kernels [15].

2.2.3. Siamese Content-Attentive GCN (SCA-GCN)

In this research, inspired by GCN and the idea of siamese network [16], our complete SCA-GCN architecture is shown in Figure 1. During forward stage, both subgraph $\mathcal{G}_{ih}$ and $\mathcal{G}_{il}$ pass through identical GCN layers $\boldsymbol{G}$ to obtain the output node (each stimuli) representation $\boldsymbol{n_{ih}}$ and $\boldsymbol{n_{il}}$. Then, self-attention mechanism [17] are utilized here for node-wise aggregation into single graph-level physiological representation. Then, by extracting the original video stimuli content vector $\boldsymbol{s_c} = \{s^1, \ldots, s^{n_c}\} \subset \mathbb{R}^{d_s}$ where d_s is the dimension of content vector, we learn a modified visual content attention $\boldsymbol{\alpha_{ic}} = \{\alpha_{ic}^k, k = 1, 2, \ldots n_c\} \subset \mathbb{R}^1$ that is calculated as:

$$\alpha_{ic}^k = \frac{exp(\boldsymbol{A}(concat[n_{ic}^k, s_c^k]))}{\sum_{k=1}^{n_c} exp(\boldsymbol{A}(concat[n_{ic}^k, s_c^k]))} \tag{2}$$

where $\boldsymbol{A}$ is a trainable network for outputting the attention weights. Note that during the calculation of the $\boldsymbol{\alpha_{ic}}$, $\boldsymbol{s_c}$ is concatenated with the corresponding node vector to obtain the attention weight. We consider this step as an injection of visual stimuli affective content in regularized learning of a graph representation for personality classification.

In this research, the visual content vector of each emotion stimuli video is extracted using the pre-trained Inception and PCA model proposed in [18] and results in image-level descriptors of dimension 1024. To prevent the curse of dimensionality, another PCA was applied over the dataset that reduces the dimension to 32, and a video-level content vectors $\boldsymbol{s_c}$ are further aggregated using mean pooling. Finally, we obtain the subgraph embedding z_{ic} as:

$$z_{ic} = \boldsymbol{n_{ic}^\intercal \alpha_{ic}} \tag{3}$$

Both z_{ih} and z_{il} are concatenated and fed into the prediction network $\boldsymbol{P}$ for persoanlity prediction. The network udpate criteria would be standard cross entropy loss.

3. EXPERIMENTAL SETUP AND RESULT

3.1. Experimental Setup

The exact architecture of our content-attentive GCN includes three blocks of networks: GCN block $\boldsymbol{G}$ consist of standard GCN layer with dimension $[488 - 24]$; attention block $\boldsymbol{A}$ is composed of a single trainable matrix with dimension $[54 - 1]$ to output single attention weight for each node; the final prediction layer $\boldsymbol{P}$ is constructed using dense layer with dimension $[48 - 2]$. Several hyperparameters were grid-searched: dropout rate between $[0.2, 0.5]$, learning rate among $[0.01, 0.005, 0.001]$. Batchsize is fixed as 16, the max epoch is 200, and the optimizer is Adam. To prevent overfitting, we carry out all experiments using a subject independent 10-fold cross-validation and report the average results of 10 independent experiments. The final evaluation metric used is the unweighted average recall (UAR).

3.1.1. Comparison Models

We first conduct our experiments utilizing linear SVM and vanilla DNN without considering structural relationships of physiological responses across video stimulus. Then we compare our methods with the following models to examine the effectiveness of our proposed content-attentive GCN:

- **Attention Multiple Instance Learning (AMIL)** [19]: Multiple instance learning has been a variation of supervised learning which is designed to predict a single label while a bag of instances is given. In this scenario, the bag could be viewed as the stimulated physiological responses while the label refers to a subject's personality. The improved AMIL adopts the self-attention mechanism as a soft-voting scheme during the bag prediction. However, comparing with graphical models, this method focuses on an instance-level prediction but omits the potential structural information between instances.

- **One-way Content-Attentive GCN (G-1-x)** The vanilla 1 way GCN models. Here 1-way refers that there would be only one graph built utilizing all of the physiological responses of a subject (instead of 2 subgraphs). There are several variations: $\boldsymbol{a}$: average all the nodes without

Table 3. A summary of average attention weights along 16 video stimulus. The bold part refers to weights larger than 0.2. '*': The highest among all videos.

Personality	Model	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Open	**G-2-b**	0.1	0.12	0.12	0.15	0.14	0.13	0.1	0.13	0.11	**0.22***	0.11	0.12	0.15	0.12	0.17	0.14
	G-2-d	0.09	0.1	0.03	0.12	**0.72***	0.05	0.11	0.04	0.02	0.06	0.04	0.09	**0.36**	0.09	0.12	0.1
Emot	**G-2-b**	0.11	0.13	0.12	0.14	0.15	0.14	0.12	0.15	0.11	0.19*	0.12	0.11	0.12	0.15	0.14	0.11
	G-2-d	0.11	0.09	0.07	0.08	**0.43***	0.1	0.02	0.19	0.09	0.16	0.12	0.15	0.1	0.14	0.14	0.12
Extr	**G-2-b**	0.12	0.13	0.1	0.16	0.12	0.12	0.1	0.14	0.09	**0.26***	0.13	0.13	0.13	0.14	0.12	0.14
	G-2-d	0.08	0.09	0.09	0.06	**0.33***	0.06	**0.29**	0.15	0.11	0.11	0.12	0.12	**0.22**	0.07	0.12	0.13

attention mechanism; $\boldsymbol{b}$: attention weights $\boldsymbol{\alpha}$ are calculated purely from self-attention; $\boldsymbol{c}$: $\boldsymbol{\alpha}$ are calculated using one-hot encoding of original video stimulation's ID as the content vector; $\boldsymbol{d}$: $\boldsymbol{\alpha}$ are calculated with visual content embedding described in section 2.2.3.

- **Siamese Content-Attentive GCN (G-2-x)** To characterize the physiological responses under different emotional inducement levels (*high/low* of either arousal or valence), we split a subject's physiological responses into two subgraphs $\mathcal{G}_{ih}, \mathcal{G}_{il}$. Then the physiological embedding with similar stimulation would be aggregated through this siamese network $\boldsymbol{G}$ into a single vector for further classification. The rest of the variations are the same as one-way GCN.

3.2. Personality Recognition Results

Table 2 summarizes our personality recognition results. Our proposed SCA-GCN(G-2-d) reaches the best UAR on Open, Emot, and Extr, which is a relative gain of 20.4%, 9% and 13.9% over the naive DNN voting technique. Several notable observations can be summarized. Firstly, using the graph to represent structural information outperforms other models for this task of personality recognition using physiology. We found that directly concatenating physiological features of a subject across all emotion stimuli would lead to an extremely large feature dimension creating issues of overfitting. As for the video-wise majority-voting technique, lack of proper integrative modeling between different responses deteriorates the personality recognition performances.

We observe that while AMIL is a strong baseline, its lack of explicit modeling on the inter-responses structure leads to unsatisfying recognition results. Furthermore, we observe for our proposed model, in dimensions of Crea, Emot, and Extr, our G-2-d outperform both G-2-c and G-2-b. This results in intriguing showing that not only knowing '*WHICH*' specific emotion stimuli could correlate with the physiological status in terms of performing personality recognition, but also '*WHAT*' the content of the stimuli itself is more important. Lastly, two-way siamese architecture enhances the model capability. During the subgraph splitting step, we manually force each subgraph could only bind edges among nodes from a similar intended emotional stimulation. We believe that it acts as a hard constraint as if we force our model to focus on learning subtle structural information of physiology under similar affective stimulation, and this fine-grained representation improves our personality recognition results.

3.3. Analysis on Attention

In this section, to understand the potential modulation of visual content stimuli toward physiological personality recognition, we gather our model's visual content-attention weights $\boldsymbol{\alpha}$ for each subject then average them into a video-level statistics (table 3). Here we only report numbers from Crea, Emot, and Extr due to their high recognition performances. We immediately observe that after incorporating the visual content vectors for attention learning, attention weights tend to be more concentrated on a smaller subset of video stimuli. For example, in Open, the key physiology shifts from video 10 toward video 5 and 13 and especially focusing on 5 with an attention weight of 0.72. A similar phenomenon is also be seen in the other two personality attributes. Also, we notice that after the embedding of the visual content, physiological responses from video 5 are consistently selected in recognizing these three personality attributes, and video 13 is especially critical for Open and Extra. It is interesting to observe that there truly exist interrelationships between personality traits and physiology which is conditioned on the content of the external media stimuli. The reason why video 5 and 13 are especially relevant will require a further detailed investigation.

4. CONCLUSION

In this work, we present a novel framework of the siamese content-attentive graph convolution network for personality recognition using physiology. The experiments show that our method reaches the known state-of-the-art personality recognition results on the AMIGO database, and the analysis reveals that the inclusion of visual content regularizes the learning to obtain a more discriminative physiological representation. To our best knowledge, this is one of the first work on APR that jointly considers the physiology (stimulated response) and the actual video content (stimulation material). There are multiple future directions. An immediate one would be verifying our results on similar datasets. Second, we will include additional modalities in the emotion stimuli, e.g., the acoustic sound of the video. By better understand exactly *what* components within a media clip that would trigger physiological responses linked to a subject's own personality would help in advancing a variety of human-centered multimedia applications [20, 21].

5. REFERENCES

[1] Alice Hall, "Audience personality and the selection of media and media genres," *Media Psychology*, vol. 7, no. 4, pp. 377–398, 2005.

[2] Tomas Chamorro-Premuzic and Adrian Furnham, "Personality and music: Can traits explain how people use music in everyday life?," *British Journal of Psychology*, vol. 98, no. 2, pp. 175–185, 2007.

[3] Marco Cristani, Alessandro Vinciarelli, Cristina Segalin, and Alessandro Perina, "Unveiling the multimedia unconscious: Implicit cognitive processes and multimedia content analysis," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 213–222.

[4] Jacob B Hirsh, Sonia K Kang, and Galen V Bodenhausen, "Personalized persuasion: Tailoring persuasive appeals to recipients personality traits," *Psychological science*, vol. 23, no. 6, pp. 578–581, 2012.

[5] Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 74–79, 2017.

[6] Golnoosh Farnadi, Geetha Sitaraman, Shanu Sushmita, Fabio Celli, Michal Kosinski, David Stillwell, Sergio Davalos, Marie-Francine Moens, and Martine De Cock, "Computational personality recognition in social media," *User modeling and user-adapted interaction*, vol. 26, no. 2-3, pp. 109–142, 2016.

[7] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L Vieriu, Stefan Winkler, and Nicu Sebe, "Ascertain: Emotion and personality recognition using commercial sensors," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, 2016.

[8] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *arXiv preprint arXiv:1702.02510*, 2017.

[9] Mojtaba Khomami Abadi, Juan Abdón Miranda Correa, Julia Wache, Heng Yang, Ioannis Patras, and Nicu Sebe, "Inference of personality traits and affect schedule by analysis of spontaneous reactions to affective videos," vol. 1, pp. 1–8, 2015.

[10] Julia Wache, Ramanathan Subramanian, Mojtaba Khomami Abadi, Radu-Laurentiu Vieriu, Nicu Sebe, and Stefan Winkler, "Implicit user-centric personality recognition based on physiological responses to emotional videos," pp. 239–246, 2015.

[11] Marco Perugini and Lisa Di Blas, "Analyzing personality related adjectives from an eticemic perspective: the big five marker scales (bfms) and the italian ab5c taxonomy," *Big Five Assessment*, pp. 281–304, 2002.

[12] Makowski. D, "Neurokit: A python toolbox for statistics and neurophysiological signal processing (eeg, eda, ecg, emg...)," Day, 01 November 2016, Paris, France.

[13] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[14] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu, "A comprehensive survey on graph neural networks," *arXiv preprint arXiv:1901.00596*, 2019.

[15] Minjie Wang, Lingfan Yu, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, et al., "Deep graph library: Towards efficient and scalable deep learning on graphs," *arXiv preprint arXiv:1909.01315*, 2019.

[16] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah, "Signature verification using a" siamese" time delay neural network," in *Advances in neural information processing systems*, 1994, pp. 737–744.

[17] Jianpeng Cheng, Li Dong, and Mirella Lapata, "Long short-term memory-networks for machine reading," *arXiv preprint arXiv:1601.06733*, 2016.

[18] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[19] Maximilian Ilse, Jakub M Tomczak, and Max Welling, "Attention-based deep multiple instance learning," *arXiv preprint arXiv:1802.04712*, 2018.

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

[21] Daniel Bone, Chi-Chun Lee, Theodora Chaspari, James Gibson, and Shrikanth Narayanan, "Signal processing and machine learning for mental health research and clinical applications [perspectives]," *IEEE Signal Processing Magazine*, vol. 34, no. 5, pp. 196–195, 2017.