MMANN: Multimodal Multilevel Attention Neural Network for Horror Clip Detection

Xingliang Cheng, Xiaotong Zhang, Mingxing Xu and Thomas Fang Zheng*
Center for Speech and Language Technologies, Research Institute of Information Technology,
Department of Computer Science and Technology, Tsinghua University, Beijing, China
*Corresponding Author E-mail: fzheng@tsinghua.edu.cn

***Abstract*—Recently, the number of the online videos is booming. However, its openness gives the horror clips a chance to threaten children's physical and mental health. Therefore, it is necessary to design an algorithm to filter the horror clips in online videos. In this paper, we proposed a multimodal multilevel attention neural network for horror clip detection. Information from visual modality and auditory modality is used to describe the various factor of horror, including violence, bloody, deformed human, scream, sudden sound, etc. The temporal-level attention is designed to give the model the ability to capture horror moments. The modal-level attention automatically balances the weight on all modalities. We evaluate the model on the same dataset used in MediaEval 2017 Emotional Impact of Movies Task. The experimental result shows the advantages of our proposed model compared with other groups.**

I. Introduction

Nowadays, the Internet has become increasingly important in daily lives. The online videos have brought us convenience. Meanwhile, we also noticed that some videos are not appropriate for all users. A study shows that horror information may translate into phobias on children [1], which means that the horror clips may hurt children's mental health. So it is necessary to classifying and filtering those undesirable content.

Horror clip detection is to find the horror clip on the entire video. Few studies have been done in this area. Wang et al. used the visual, audio and color emotion features to describe the horror [2], and recognized the horror video scene via multiple-instance learning [3]. Ding et al. [4] proposed a multi-view multi-instance learning model through sparse coding to integrating the context cues among instances.

Although little researches have been done in this field, the research work on video affective content analysis is flourishing. Video affective content analysis aims to automatically recognize the emotion induced by video, it could reduce to a horror clip detection problem when the target emotion is limited to horror. In this field, Jean H [5] used sound energy and object motion as the low-level feature to identify the slapstick comedy. TS et al. [6] proposed a hybrid SVM-RBM classifier to recognize the emotion in the video. Soujanya et al. [7] fused the visual, audio and textual modality for sentiment analysis.

There are two main challenges in horror clip detection. First, the factor that induces fear is various, including violence, bloody, deformed human, scream, sudden sound, etc. So it's difficult to describe horror. Second, even in the most horrifying movies, not all the moments are scaring. Only a fraction of the time can be horror, which we called "the horror moment". However, the existing database may not have the exact time of horror moment, so the model needs to find it by itself.

In this paper, we propose the multimodal multilevel attention neural network model for horror clip detection. To better describe horror, information from multimodality is used, and several streams of the feature vector are extracted from each modality. Temporal-level attention is designed to capture the horror moment automatically. Modal-level attention can focus on the more important modality.

The remainder of the paper is organized as follows. Section II introduces the related work. Section III describes the entire horror clip detection framework. The proposed model, which is the most important part of the whole framework, is described in Section IV. The experimental configuration is described in Section V, and the result is presented in Section VI. Section VII concludes the paper with a brief summary and description of future work.

II. Related Work

Resently, deep neural network has been used in video affective content analysis. Saowaluk C et al. [8] proposed a unique sieving-structured neural network to classify movie clips into three type of emotion. Lei et al. [9] and Quan et al. [10] built a multimodal deep regression Bayesian network to capture the relation between the modalities for affective video content analysis.

As an effective method, attention mechanism has been widely used in many fields. Mnih et al. [11] introduced a visual attention model that selecting a sequence of location by attention mechanism and only processing those regions with high resolution. It uses the hard attention mechanism, which is non-differentiable, so the model needs to be trained using reinforcement learning methods. Bahdanau et al. [12] introduced the soft attention mechanism for neural machine translation, which reveals the alignments between the source and target sentence. Unlike hard attention, soft attention is differentiable, so it can be trained easily. Xu et al. [13] introduced an attention-based model for image caption generation. It can automatically learn the latent alignment between caption and image. Zichao et al. [14] proposed a hierarchical attention network for document classification. The network has a word-level attention layer followed by a sentence-level attention

layer, which allows the model to learn step by step.

III. Horror Clip Detection Framework

Horror clip detection aimed at giving an alarm when something horror appears. To ensure the sensitivity, the timescale of detection result should be small. But emotion is not a short-term phenomenon, the result based on a short clip may not be reliable. So we recognize the horror in the relatively long clips, and then transform the result into small timescale detection result. The overall framework is shown in Fig. 1. Each part described below:

- **Window Sliding.** The whole movie is cut into overlapping clips. Each clip has an appropriate length (15-second in this paper).
- **Horror Clip Classification.** Each clip is judged that whether it contains any horror moment. The multimodal multilevel attention neural network (MMANN) is used for classification and will be described in Section IV.
- **Detection Result Processing.** The detection result is generated based on the classification result of overlapping clips. In this paper, the timescale of the detection result is one second. Because of overlapping, each second of the video is contained by more than one clip. So we simply select the maximum possibility of those clips as the horror possibility at this second.

In addition to detection, a common task that requires determining whether a clip from a long movie is horror or not. There are two ways to do it:

- Use the horror clip classification model directly. In this way, the length mismatch of clips during training and testing can lead to problems.
- Make the decision based on the detection result. In this way, we can handle the clips with any length, and make full use of the context information of the movie.

In this paper, we choose the second way to do it. The horror possibility of one clip equals the maximum horror possibility of all the seconds this clip contained. If it is greater than 0.5, then this clip is classified as horror, and vice versa.

IV. Multimodal Multilevel Attention Neural Network

Horror clip classification is the most important part of the whole framework. Its structure is shown in Fig. 2. First, for a movie clip, several streams of feature vector are extracted from visual and audio modality. Then, all the feature streams are synchronized by dividing them into the same overlapping segments, each segment is processed by LSTM to learn a high-level representation. Next, temporal-level attention is performed in segment representations. After that, the output is further merged with modal-level attention to form the embedding of this video clip. Finally, a decision (horror or non-horror) is made.

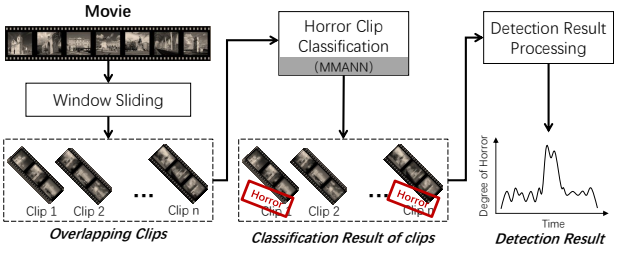

Fig. 1. The illustration of horror clip detection framework.

A. Feature Extraction

Visual modality is the main source for human to obtain information, it's also the main cause of horror. Some important factor of fear like blood and violence is mainly presented through visual modality.

There are two things important: what is appearing on the screen and how is it looks like. In view of this, two kinds of the feature are extracted from visual modality: the embedding vector of the image (called ***VGG16-fc6*** in this paper) and the information about the objects that appearing on the screen (called ***ObjInfo*** in this paper). The object information answered what's on the screen explicitly, and the embedding of the image answered both questions implicitly.

The embedding of the image is extracted from VGGNet [15]. The ObjInfo feature is constructed in the following way. Each image is processed by object detection model to detect the objects that appearing on this image. The ObjInfo feature of this image is the concatenating vector of all attribute values of all types of objects, where all the attribute values are scalar, shown as follows:

1) **Occurrence**. The number of times this type of the object appears in the image.
2) **Confidence**. The confidence score provided by the model. Choose the maximum score if more than one occurrence.
3) **Area**. The size of the object. Choose the maximum area if there is more than one occurrence.
4) **Horizontal Position**. The transverse distance between object center and the screen center. Choose the object with the maximum area if there is more than one occurrence.
5) **Vertical Position**. The longitudinal distance between object center and the upper edge of the screen. Choose the object with the maximum area if there is more than one occurrence.

In addition to the visual modality, audio also plays an important role in foiling atmosphere and expressing emotion. A horror movie will be less horror if the soundtrack is removed. Therefore, two kinds of the feature are extracted from auditory modality: the traditional MFCC and prosody feature (called ***MFCC+Prosody*** in this paper) to describe what the audio sounds like, and emotion-related feature (called ***EmoInfo*** in this paper) to describe what emotion conveyed by the audio.

Since each of the feature streams is used to describe one modality, we call it the **modal descriptor** below.

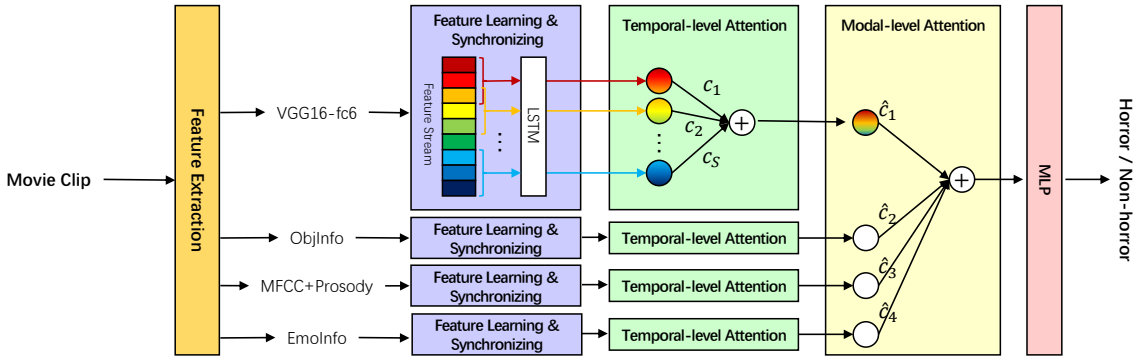

Fig. 2. Multimodal multilevel attention neural network (MMANN).

B. Feature Learning and Synchronizing

Different modal descriptors have different timescales. For example, the frame rate of the video is usually 24 fps, but the sample rate of the audio is much higher, such as 16kHz. This allows the audio to have a higher time-resolution. Besides, the timescale of modal descriptor also related to the modal descriptor itself, that is, the traditional MFCC has higher time-resolution than the statistic based emotion-related modal descriptor. That mismatch makes it hard to merge different modal descriptors later.

Therefore, we synchronize all the modal descriptors by cutting them into the same overlapping segments, as shown in Fig. 3. At each segment, LSTM is used to accept a sequence of feature to form a high-level representation (called the **segment embedding**). Notice that the LSTM shares parameters on different segments of the same modal descriptor. As for different modal descriptor, the parameters of LSTM are independent.

C. Temporal-level Attention

Even in the most horrifying movies, not all the moments are scaring. Some time is always needed to describe the environment or foil atmosphere. So it would be nice if the model has the ability to choose the most important segments which cause the fear significantly.

The temporal-level attention is designed for this purpose. For each modal descriptor, the segment embeddings are fused according to (1):

$$\boldsymbol{v}^{(m)} = \sum_{i=1}^{s} \boldsymbol{c}_i^{(m)} \boldsymbol{E}_i^{(m)} \tag{1}$$

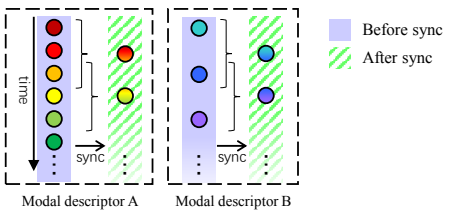

Fig. 3. An illustration of synchronizing two modal descriptors.

where $\boldsymbol{v}^{(m)}$ is the combined vector of modal descriptor m (called the **modal descriptor embedding**), s is the number of the segments, $\boldsymbol{E}_i^{(m)}$ is the i^{th} segment embedding of modal descriptor m, $\boldsymbol{c}_i^{(m)}$ is the attention weight on i^{th} segment embedding of modal descriptor m, given by (2):

$$\boldsymbol{c}^{(m)} = softmax(\boldsymbol{W} \cdot concat(\boldsymbol{E}^{(m)}) + \boldsymbol{b}) \tag{2}$$

where $concat(\cdot)$ convert a matrix to a vector. $\boldsymbol{W}$ is a weight matrix and $\boldsymbol{b}$ is the bias vector. In this way, the model has the ability to learn to output a higher attention weight on the segment which is more related to horror.

D. Modal-level Attention

As mentioned before, different modalities have different importance. So it is reasonable to have different weights. In addition to the modality itself, the weight of modality also related to the content. For scenes of blood and violence, visual modality is more important. For some dark scenes, auditory modality is more important. Because you can't see anything at this point, but you can still be scared by a sudden screaming.

The modal-level attention is designed to adjust the weight of modal descriptor according to the current video content. The calculation of modal-level attention is the same as that of the temporal-level attention, except that the $\boldsymbol{E}_i^{(m)}$ appearing in (1) and (2) represents the modal descriptor embedding now.

E. Decision Making

After previous parts, an embedding of the video clip is generated. The last thing is using this embedding to determine whether this clip is horror or not. There are many classical models, such as Support Vector Machine (SVM), Random Forest (RF), etc., can be used to make the decision. In order to train the entire model easily using the backpropagation algorithm, the Multilayer Perceptron (MLP) is chosen as the decision model. Other decision models will be explored in the future.

V. Experiments

A. Database

The model was evaluated on the same dataset used in MediaEval 2017 Emotional Impact of Movies Task [16],

TABLE I
DATA DISTRIBUTION

	#Movies	#Horror clips	#Non-horror clips
Training Set	24	251	4161
Validation Set	6	32	830
Test Set	14	204	5506

which is a derived from LIRIS-ACCEDE database [17]. The development set consists of 30 movies, with a total length of 7.37 hours. The test set consists of 14 movies, with a total length of 7.95 hours. All the movies were cut into 10-second clips, with 5-second overlap between the adjacent clips. Each clip is labeled as horror or not. Six movies in the development set were chosen as the validation set, and the remaining form the training set. Table I shows the details.

Notice that the horror clips are rare, may not be sufficient to train the model. So we recut all movies into 15-second clips with 1-second overlap to perform data augmentation. (Acutally, this is the "window sliding" part of the framework, described in Section III). Then the biggest problem is to generate the label of those clips in the training set, because the original label is based on 10-second clips with 5-second overlap. There is a fact that every horror clip must contain one horror moment at least. That means the non-horror clip doesn't contain any horror moment. So for each 15-second clip, if it contains an entire 10-second clip which is labeled as horror, then this 15-second clip is horror too. If it is covered by consecutive non-horror 10-second clips, it's non-horror too. Otherwise, the label of 15-second clip stays unknown, since we don't know the precise time of horror moment. For simplicity, all the unknown clips in the training set were discarded.

Notice that the movie after data augmentation is the input of MMANN, which is a model used in "horror clip classification" part. In test stage, the output of MMANN will be further processed to form the result of origin movie clips (described in Section III), that means the test set is strictly same as the dataset used in MediaEval 2017.

B. Feature Extraction Configuration

For visual features, the video was first sampled into an image stream. The embedding vector of the image was extracted from VGGNet (VGG16 [15] fc6 layer)[1] every one second, provided by MediaEval 2017 organizer [16], using Matlab Neural Networks toolbox[2]. The object appearing on the image was detected by faster R-CNN with inception resnet v2 model [18] (pre-trained on the open image dataset [19]). The detection result includes the object's name, the confidence and the box location of the object. All the types of objects detected in the training set were filtered by the mutual information between the type of the object and whether it contained by a horror clip. Only top 25 types of objects were selected, and the rest was ignored. We described each type of objects with 5 attributes, so the dimension of ObjInfo feature is $5 \times 25 = 125$. Two images per second were sampled from video to obtain the ObjInfo stream.

For audio features, traditional MFCC and prosody were extracted by openSMILE [20] toolbox with 20ms window length and 10ms window shift. To make training easier, those features were downsampled to represent 1-second length audio segment with 0.5-second shift, by averaging all the features in the downsampling window. The emotion-related feature was extracted by openSMILE [20] toolbox for every 5-second audio segment with 1-second shift. The default configuration named "emobase2010.conf" was used.

C. Experimental Configuration

In the experiment, the hyperparameter was tuned according to the results of cross-validation. The final parameters are described below. The number of LSTM node was 64. The MLP had one hidden layer with 64 nodes using the ReLU activation function and an output layer with 1 node using the sigmoid activation function. All parameters were initialized by Xavier normal initializer [21].

Binary cross-entropy was used in loss function with regularization terms. L2 regularization was performed on all parameters with the weight of 1. In order to reduce overfitting, it was also performed on the modal-level attention weights with the weight of 0.0005. Moreover, dropout technique [22] was used before and between LSTM nodes with a ratio of 50%.

The model was optimized by Adam algorithm [23] with learning rate of 0.0005. The batch size was 256. At each epoch, an equal number of positive and negative clips were randomly selected from the training set. Considering the amount of training data is limited, it may cause the overfitting problem. We stopped training after 5 epochs according to the loss evaluated on the validation set.

D. Performance Evaluation Metrics

The model was evaluated by precision, recall, and f1.[3] All the metric are averaged over all movies, as showed in (3):

$$\hat{M} = \frac{1}{\text{\#Movie}} \sum_{i=1}^{\text{\#Movie}} M_i \tag{3}$$

where the $\hat{M}$ means a movie-level metric, could be precision, recall or f1. M_i means the corresponding metric calculated in movie i.

Due to the randomness of the model, the result of our model shown below is the average result of five independent trials.

VI. RESULTS AND DISCUSSION

A. Contribution of Attention Mechanism

In order to analyze the contribution of attention mechanism, we disable some of them by fixing the attention weights all the same. As shown in Table II, the precision is low when both attention mechanisms are disabled. When there is only

[1] https://www.mathworks.com/help/nnet/ref/vgg16.html
[2] https://www.mathworks.com/products/neural-network.html
[3] We don't use accuracy as a metric due to the high imbalance of data.

TABLE II
CONTRIBUTION OF THE ATTENTION MECHANISM

	Precision	Recall	F1
No Attention	0.2393	**0.6570**	0.2880
Temporal-level Attention (TLA)	**0.2682**	0.6228	0.2966
Modal-level Attention (MLA)	0.2521	0.6202	0.2899
TLA + MLA	0.2562	0.6511	**0.3106**

Note: *The maximum value of all metrics is* ***0.7143*** *($\frac{14-4}{14}$), because there are four movies which don't contain any horror moment.*

one kind of attention mechanism, temporal-level attention is more effective than modal-level attention. From the f1 metric, it can be seen that when both attention mechanisms are used together, the overall performance of the model is improved.

Comparing the last three rows, the recall is lower when some attention mechanism is disabled. This may be related to the way to disable the attention mechanism. That is, the average weighting on temporal-level may conceal the horror moment and the average weighting on modal-level makes the important modality overshadowed by other less relevant modalities. So the recall raises when both levels enable the attention mechanism. As for why the recall is so high when all attention mechanism is disabled, more research is needed in the future.

B. Contribution of Modal Descriptors

To explore the contribution of the different modal descriptor, Table III shows the result of using only some of the modal descriptor. The first four rows of the table show the results when only one modal descriptor is used. In this case, VGG16-fc6 works best. Next four rows of the table show the results of using three modal descriptors. Performance drops the most when there is no VGG16-fc6, followed by no ObjInfo. Next two rows show the results of using modal descriptors belong to visual or auditory domain only, the performance of using only visual domain modal descriptors are very close to the best performance.

All the results above show that the modal descriptors from visual domain contribute much more than those from auditory domain, and the VGG16-fc6 contributes most. The last line of the table shows the result when all modal descriptors are used, it further improves overall performance, according to f1.

C. Performance Comparison

Table IV shows the result of our model compares with other groups who take part in MediaEval 2017 Emotional Impact of Movies Task [16]. Our model outperforms three of them on all three metrics. For MIC-TJU, our model's precision is lower, but both recall and f1 are much higher. As an assistant tool to help humans to filter horror clips, the value of recall determines the quality of work, and the value of precision determines the quantity of work. So those two kinds of the model can work together in the actual system.

D. Discussion

Visual analysis is a good way to understand how the model works. We select a movie from validation set and show how

TABLE III
CONTRIBUTION OF MODAL DESCRIPTORS

Visual		Auditory		Precision	Recall	F1
VGG	Obj	M+P	Emo			
✓				0.1819	0.6321	0.2421
	✓			0.1646	0.3839	0.1623
		✓		0.1278	0.5124	0.1682
			✓	0.1346	0.4600	0.1798
	✓	✓	✓	0.1170	0.4040	0.1406
✓		✓	✓	0.2201	0.6506	0.2796
✓	✓		✓	0.2524	0.5918	0.2947
✓	✓	✓		**0.2678**	0.6113	0.3045
✓	✓			0.2518	0.6285	0.3073
		✓	✓	0.1277	0.4911	0.1631
✓	✓	✓	✓	0.2562	**0.6511**	**0.3106**

Note: ***VGG****=VGG16-fc6;* ***Obj****=ObjInfo;* ***M+P****=MFCC+Prosody;* ***Emo****=EmoInfo.*

TABLE IV
COMPARISON WITH MEDIAEVAL 2017 RELATED WORKS

	Precision	Recall	F1
HKBU [24]	0.1688	0.0657	0.0786
TCNJ-CS [25]	0.2553	0.1922	0.1740
THUHCSI [26]	0.2318	0.2781	0.2352
MIC-TJU [27]	**0.3756**	0.0991	0.1424
MMANN	0.2562	**0.6511**	**0.3106**

the modal-level attention weights vary over time, as shown in Fig. 4. The vertical axis represents different modal descriptors, and the horizontal axis represents the overlapping clips in chronological order. We find that model pays more attention to VGG16-fc6, and it does contribute the most, according to Table III. That means the model does learn the different importance of modal descriptors.

We also observed the temporal-level attention weights, but they do not change much over time and the pattern of change is not significant. This may due to the small number of horror clips which are not enough to support the model to learn the pattern of horror moment when there is no precise labeling of it. Besides, the cause of fear is variety, making the training data of the same pattern more scarce. However, considering that temporal-level attention is still helpful according to Table

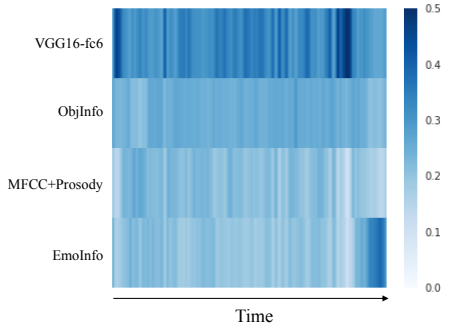

Fig. 4. An example of modal-level attention weights varies with time.

II, we will continue to study and refine it in the future.

VII. Conclusions

In this paper, we proposed the multimodal multilevel attention neural network for horror clip detection. Multimodal aims to better describe the factor which induces fear. Temporal-level attention mechanism is designed to capture the horror moment. Modal-level attention mechanism automatically balances the weight on all modalities. We evaluate the model on the same dataset used in MediaEval 2017 Emotional Impact of Movies Task. Our model shows a huge advantage on recall and outperforms all other groups on f1.

In the future, we will try to extract more features, such as the semantic information or the type of the movie, to better detect the horror content. In addition, we will analyze the reason for the low precision and try to propose some strategies to improve it. Furthermore, this model could be adapted to other types of content, for instance, the porn scenes. By now, just horror scenes were evaluated, more scenes will be evaluated and explored in the future.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China under Grant No. 61433018 / 61171116 / 61633013.

References

[1] A. P. Field and J. Lawson, "Fear information and the development of fears during childhood: Effects on implicit fear responses and behavioural avoidance," *Behaviour Research and Therapy*, vol. 41, no. 11, pp. 1277–1293, 2003.

[2] J. Wang, B. Li, W. Hu, and O. Wu, "Horror movie scene recognition based on emotional perception," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 1489–1492.

[3] J. Wang, B. Li, W. Hu, and O. Wu, "Horror video scene recognition via multiple-instance learning," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 1325–1328.

[4] X. Ding, B. Li, W. Hu, W. Xiong, and Z. Wang, "Horror video scene recognition based on multi-view multi-instance learning," in *Asian Conference on Computer Vision*. Springer, 2012, pp. 599–610.

[5] J. H. French, "Automatic affective video indexing: Sound energy and object motion correlation discovery," in *Southeastcon, 2012 Proceedings of IEEE*. IEEE, 2012, pp. 1–6.

[6] T. Ashwin, S. Saran, and G. R. M. Reddy, "Video affective content analysis based on multimodal features using a novel hybrid svm-rbm classifier," in *Electrical, Computer and Electronics Engineering (UPCON), 2016 IEEE Uttar Pradesh Section International Conference on*. IEEE, 2016, pp. 416–421.

[7] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, 2016.

[8] S. C. Watanapa, B. Thipakorn, and N. Charoenkitkarn, "A sieving ANN for emotion-based movie clip classification," *IEICE transactions on information and systems*, vol. 91, no. 5, pp. 1562–1572, 2008.

[9] L. Pang and C.-W. Ngo, "Mutlimodal learning with deep boltzmann machine for emotion prediction in user generated videos," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015, pp. 619–622.

[10] Q. Gan, S. Wang, L. Hao, and Q. Ji, "A multimodal deep regression bayesian network for affective video content analyses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5113–5122.

[11] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

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

[13] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Samullakhudinov *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048–2057.

[14] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] E. Dellandréa, , M. Huigsloot, L. Chen, Y. Baveye, and M. Sjöberg, "The MediaEval 2017 emotional impact of movies task," in *Working Notes Proceedings of the MediaEval 2017 Workshop*, Dublin, Ireland, 2017.

[17] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.

[18] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning." in *AAAI*, vol. 4, 2017, p. 12.

[19] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova *et al.*, "Openimages: A public dataset for large-scale multi-label and multi-class image classification." *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017.

[20] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[21] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] Y. Liu, Z. Gu, and T. H. Ko, "HKBU at MediaEval 2017 emotional impact of movies task," in *Working Notes Proceedings of the MediaEval 2017 Workshop*, Dublin, Ireland, 2017.

[25] S. Yoon, "TCNJ-CS@ MediaEval 2017 emotional impact of movie task," in *Working Notes Proceedings of the MediaEval 2017 Workshop*, Dublin, Ireland, 2017.

[26] Z. Jin, Y. Yao, Y. Ma, and M. Xu, "THUHCSI in MediaEval 2017 emotional impact of movies task," in *Working Notes Proceedings of the MediaEval 2017 Workshop*, Dublin, Ireland, 2017.

[27] Y. Yi, H. Wang, and J. Wei, "MIC-TJU in MediaEval 2017 emotional impact of movies task," in *Working Notes Proceedings of the MediaEval 2017 Workshop*, Dublin, Ireland, 2017.