Original Research Paper

Volume - 12 | Issue - 02 | February - 2022 | PRINT ISSN No. 2249 - 555X | DOI : 10.36106/ijar

Computer Science

MULTIMODAL ANALYSIS OF VIDEO WITH NEURAL NETWORKS

Akshat Vedant* Computer Science, PES University, Bengaluru. *Corresponding Author

ABSTRACT The cyberspace technology in age of social-media, has introduce an innovative and ravishing way of passing the information in form of online streaming videos. Interestingly, videos are able to communicate a complex message in some seconds of duration involving all auditory and visual senses. Hence, somewhere it is needed to monitor the videos, its shared context and its affect over viewers. Thereby, it becomes imperative to test the negative, positive and neutral sentiments hidden within the video which is also termed as sentiment analysis of videos. For this text, audio and visuals are known modalities in consideration to relation with utterances. This paper presents the contextual multimodal sentiment analysis using the technique of neural network using attention mechanism.

KEYWORDS :

INTRODUCTION

Every day a huge amount of information are being shared over the social-media platforms like Facebook, YouTube, Vimeo in format of videos. A video comprised of information channels known as modalities which are text modality, audio modality and visual modality. Thus, the need of mining opinions from such a large quantity of videos calls a popular field of research called multimodal sentiment analysis **[1]**. There exist various approaches to this which consider each utterance as independent entities. This paper proposes Long Short-Term Memory Network (LSTM) network in contextual relationship among the utterances for classifying the target utterance.

Problem Definition

Let us consider a video defined in terms of utterances as: $V_n = [u_{n1}, u_{n2}, ..., u_{nm}, ... u_nN_n]$. Where, u_{nm} is m^{th} utterance in given video and u_nN_n is total number of utterances in the video. Every utterance u_{nm} *are* labels speaker's sentiment. Henceforth, classification of utterance u_{nm} in video is dependent over other utterances calculated as $[u_{nx} \mid \forall x \leq N_{nx} \neq m]$.

Utterance Feature Extractions

A. Unimodal Features Extraction

Any of the single modalities and not any contextual relationships are considered in this technique which are denoted as unimodal classifiers. This is of following types: -

Textual Features Extractions

A video has utterances of interleaved convolution layers in window of 2x2. Each layer contains kernels and feature maps.

First Layer [Size (3,4), Feature Maps = 50] and Second Layer [Size (2), Feature Maps = 100] with fully connected layer of size 500. With activation factor *ReLU,* it produces V_{max} output considers as text modality utterance.

Audio Features Extractions

The openSMILE 3.0 toolkit helps in audio extraction with frame-rate of 30 *Hz* or more and a sliding window of 100 *ms* along with descriptors namely voice pitch, intensity, and their statistical mean, quadratic mean and root.

Visual Features Extractions

The videos have frames of changing images chained up in a sequence **[12][13]**. The extractions of which are performed with three-dimensional Convolutional Neural Network (3D-CNN) technique. For this, let a video V in relation to utterance video represented as $V \in U^{(ch*f*h*w)}$. Here ch is number of channels in an image. The estimated valued used here for ch = 3 for RGB mode of image classification. f is number of frames in video, h is height of each frame in video and w is width of each frame in video. This originality of video V is then applied with 3D-CNN convolutional filter F represented as $U^{(fm*ch*fd*fh*fw)}$. Here, fm is total number of feature maps of filter, fd is total number of frames of filter, fh is height of the filter and fw is width of the filter. This filter F thus slides video V and produces output represented as $V_{out} \in R^{(fm*ch*(f-fd+1)(h-fh+1)(w-fw+1))}$. Irrelevant features are discarded over the window dimension 3 × 3 × 3 on vout and connected to layer size 280, followed by a Vmax layer for furthermore classifications of video V.

B. Multimodal Sentiment Analysis

AT-Fusion: Attention Based Network - Multimodal Fusion

An attention network or AT-Fusion network takes inputs out of text, audio and visual modalities and results an attention score for each modality. Attention mechanism is used in image classification **[14]** to focus on parts of an object relevant for classification of the deep neural networks. These dimensions of the feature vectors of all three modalities are fed into attention network using a fully connected layer of size *s* in Unimodal-SVM to produce output. This output F is further provided to the CAT-LSTM (*Figure 1*) for final multimodal classification. Let V be the proposed input video with following feature sets of dimension vectors [Va, Vv, Vt] scaled on layer of size *s*, where Va is acoustic feature, is visual feature, and Vt is textual features; Thus, the fused multimodal feature vector F, input video $V \in U^{s*3}$ and attention weight Vv vector α_{fusion} are computed as:

$$P_F = \tanh(L_F.V)$$
$$\alpha_{fusion} = V_{max}(\omega_F^T.P_F)$$
$$F = V.\alpha_{fusion}^T \quad \text{..... Eq. 1.}$$

CAT-LSTM – Attention based LSTM model

In this approach, the attention-based LSTM network output obtained from *Eq.1* is given as an input of sequence of utterances per video. This generates a new representation of utterances based on the learning set of utterances **[15]**. This LSTM is a kind of specialized Recurrent Neural Network (RNN) that models long range dependencies in a sequence. Say $y \in U^{(s*M)}$ *to* be an input to the LSTM network, where M is the number of utterances in a video. The matrix y is represented as $y = [y_1, y_2, ..., y_t, ...y_M]$, *where* y_t for $t = 0$ *to* M. Each matrix cell in $\in U^s$ LSTM is computed as:

$$c_t = f_t \odot c_{t-1} + m_t \odot \tanh(W_c.y + v_c)$$
$$h_t = o_t \odot \tanh(c_t).$$

Here, Learned Training Parameter Sets are $L_m, L_f, L_o \in U^{s*sd}$ and $v_m, v_f, v_o \in U^s$. $\odot$ is element-wise multiplication. The output of this LSTM layer is represented as $H \in U^{s*M}$, where $H = [h_1, h_2, ..., h_s, ..., h_M]$ and $h_m \in U^s$ Where each modified LSTM cell output $\mathbf{h}^t$ is sent into a $\mathbf{V}_{max}$ layer for classification.

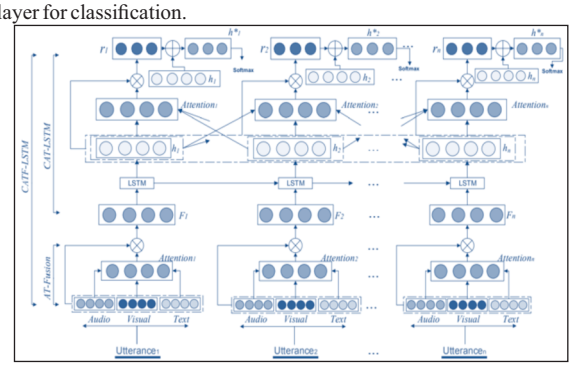

Figure 1: Output of multiple modalities from At-Fusion is fed as Input to CAT-LSTM getting used in classification.

Training Of Multimodal Classification

The Output of multiple modalities from At-Fusion is fed as Input to CAT-LSTM performs multimodal classification (*Figure 1*) also known as Contextual Attentive Fusion LSTM, which has been accomplished using two types of frameworks: -

Single-Level Framework

This framework is provided in *Figure 1* which fuse context-independent unimodal features to LSTM network for multimodal fusions.

Multi-Level Framework

In this experiment, videos are padded with dummy utterances to enable batch processing. This uses bit-masks to mitigate proliferation of noise in the network which are typically trained for 400 − 900 epochs with an early-stopping patience of 50 epochs. Contextual unimodal features improve performance of multimodal fusion framework for final classification. The classifiers are trained in an end-to-end manner using back propagation with given objective function of log – loss as:

$$loss = -\sum_i \sum_j \log\left(D_t[c_m^n] + \lambda\|\theta\|^2\right)$$

Where, c = target class, D_t = predicted distribution of n^{th} utterance from video V_m, where m ∈ [0, N] and n ∈ [0, N_m]. λ is the N_2 term for regularization and θ is the parameter set $\theta = \{K_m, v_m, K_f, v_f, K_o, v_o, K_t, k_t, K_h, k, K_p, K_y, K_{soft}, V_{soft}\}$.

DATA ANALYSIS AND RESULTS

A. MOSI Dataset

The videos are segmented into utterances using Zadeh et al. **[16]** multimodal sentiment analysis dataset called Multimodal Opinion-Level Sentiment Intensity (MOSI) dataset, consisting of 3234 utterances with its sentiment label namely positive and negative. The train set comprises of 78 individuals in the dataset, 51 videos by 37 speakers. The dataset is using a total of 1603 utterances in the training and 795 utterances to test the models. This LSTM has performed with 0.3 – 0.7 %.

B. AT-Fusion Performance

Table 1 presents the performance of CAT-F-LSTM for sentiment classification. AT-Fusion integrated within network variants amplifies the contribution of important modalities during fusion, it outperforms the simple fusion method (Simple-LSTM) in *Table 2* with V_{max} output in Unimodal-SVM.

Table 1: Comparison between single-level and multi-level fusion using CAT-F-LSTM network. The table reports the macro-fscore of classification. A=Audio; V=Visual; T=Textual.

Modality	Single-Level		Multi-Level	
	Feat Append	*AT-Fusion*	*Feat Append*	*AT-Fusion*
A + V	61.0	**61.6**	62.4	**62.9**
A + T	78.5	**79.2**	79.5	**80.1**
V + T	**78.3**	78.3	79.6	**79.9**
A + V + T	78.9	**79.3**	81.0	**81.3**

Table 2: The table reports the macro-fscore of classification. Note: feat-append=fusion by concatenating unimodal features. Multilevel framework is employed. A=Audio; V=Visual; T=Textual.

Modalities	MOSI Dataset					
	UNI-SVM	*Simple-LSTM*		*CAT-LSTM*		*ATS-FUSION*
	Feat-App	*Feat-App*	*AT-Fusion*	*Feat-App*	*AT-Fusion*	
A	58.1	59.5	-	60.1	-	-
V	53.4	54.9	-	55.5	-	-
T	75.5	77.2	-	79.1	-	-
A + V	58.6	61.4	61.8	62.4	62.9	59.1
A + T	75.8	78.5	79.1	79.5	80.1	76.3
V + T	76.7	78.7	79.1	79.6	79.9	77.5
A + V + T	**77.9**	**80.1**	**80.6**	**81.0**	**81.3**	78.3

C. Qualitative Analysis

Following are the observations on the learned attention parameters for both CAT-Fusion and AT-Fusion:

I). The context dependency is prime importance for utterance classification. For instance, the utterance: *She sells the seashells*. has implicit sentiment expressed forming the baseline unimodal-SVM However, information from neighboring utterances such as: 1) *And the dialogue threw me off*; 2) *The whole movie had a really dry dialogue* etc. indicates the negative context for the utterance.

ii) Considering the movie review utterance: *You never know what's going to happen.* This sentence does not provide explicit sentiment cues. *Figure 2.1* shows the attention weights across the video. Here text modalities are improved over audio and visual vectors. The most relevant utterances are U10, U1 (*Figure 2.1*). Here, the most important utterance U10 is located away from target utterance U4 proving the effectiveness in long distance sequence. *Figure 2.2* shows the contribution of each modality for the multimodal classification. Concluding, text has been given the highest weight followed by audio and visual by the *AT-fusion* network.

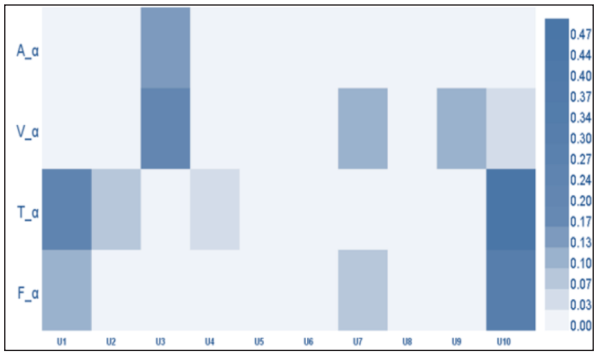

Figure 2.1: The visualization of modality scores of unimodal AT-Fusion and CAT-LSTM

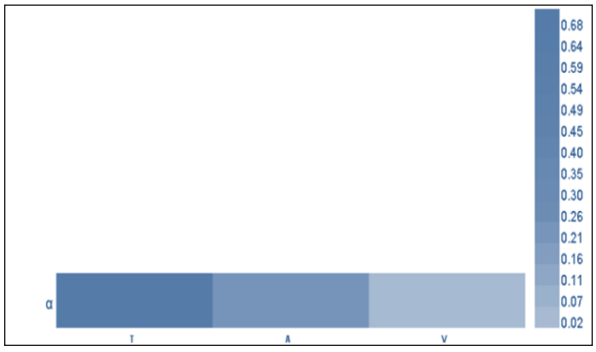

Figure 2.2: The Visualization of modality weights of the AT-Fusion network in multimodal CAT-LSTM.

CONCLUSION

Rejecting the assumption of utterance independence with contextual information obtained from the other utterances in a video while classifying target utterance. The utterances located at farthest points from the target utterances performs better.

REFERENCES

[1] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis", in *Proceedings* of EMNLP, 2015, pp. 2539–2544.
[2] Y. Kim, "Convolutional neural networks for sentence classification", *ArXiv preprint arXiv*:1408.5882, 2014.
[3] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank", in Proceedings of EMNLP, 2013, pp. 1631–1642.
[4] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush, "Character-aware neural language models", *ArXiv preprint arXiv*:1508.06615, 2015.
[5] E. Cambria, "Affective computing and sentiment analysis", IEEE Intelligent Systems, vol. 31, no. 2, pp. 102–107, 2016.
[6] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition", in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, IEEE, 2013, pp. 3687–3691.
[7] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis", in Data Mining (ICDM), 2016 IEEE 16th International Conference on, IEEE, 2016, pp. 439–448.
[8] L. C. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information", in Proceedings of ICICS, IEEE, vol. 1, 1997, pp. 397–401.
[9] L. S. Chen, T. S. Huang, T. Miyasato, and R. Nakatsu, "Multimodal human emotion/expression recognition", in Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, IEEE, 1998, pp. 366–371.
[10] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis", Journal on Multimodal User Interfaces, vol. 3, no. 1-2, pp. 33–48, 2010.
[11] B. Schuller, "Recognizing affect from linguistic information in 3d continuous space", IEEE Transactions on Affective Computing, vol. 2, no. 4, pp. 192–205, 2011.
[12] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition", IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 1, pp. 221–231, 20.
[13] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks", in Proceedings of the IEEE International

Conference on Computer Vision, 2015, pp. 4489–4497.
[14] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention.", in ICML, vol. 14, 2015, pp. 77–81.
[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory", Neural computation, vol. 9, no. 8, pp. 1735–17.
[16] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages", IEEE Intelligent Systems, vol. 31, no. 6, pp. 82–88, 2016.
[17] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis", in Proceedings of EMNLP, 2015, pp. 2539–2544.