Sequential Late Fusion Technique for Multi-modal Sentiment Analysis

Debapriya Banerjee
The University of Texas at Arlington
Arlington, Texas, USA
debapriya.banerjee2@mavs.uta.edu

Fotios Lygerakis
The University of Texas at Arlington
Arlington, Texas, USA
fotios.lygerakis@mavs.uta.edu

Fillia Makedon
The University of Texas at Arlington
Arlington, Texas, USA
makedon@uta.edu

ABSTRACT

Multi-modal sentiment analysis plays an important role for providing better interactive experiences to users. Each modality in multi-modal data can provide different viewpoints or reveal unique aspects of a user's emotional state. In this work, we use text, audio and visual modalities from MOSI dataset and we propose a novel fusion technique using a multi-head attention LSTM network. Finally, we perform a classification task and evaluate its performance.

CCS CONCEPTS

• **Computing methodologies** → **Modeling methodologies**.

KEYWORDS

late fusion, multi-modal sentiment analysis, multi head attention recurrent neural networks

ACM Reference Format:
Debapriya Banerjee, Fotios Lygerakis, and Fillia Makedon. 2021. Sequential Late Fusion Technique for Multi-modal Sentiment Analysis. In *The 14th PErvasive Technologies Related to Assistive Environments Conference (PETRA 2021), June 29-July 2, 2021, Corfu, Greece.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3453892.3461009

1 INTRODUCTION

Social media is an industry that relies on efficient emotion recognition and sentiment analysis (SA), in order to provide relevant content to the users by exploiting their preferences and habits [1]. Furthermore, the continuously increasing popularity of social robots is another eminent field that can be greatly benefited by tracking user's mental state, offering more organic and engaging human-robot interaction [8] or user-specific services [7].

2 RELATED WORK

The ever-growing user-provided content on social media consisting of different modalities, i.e., audio, text and video, has already created huge datasets [11], that eased the utilization of neural network (NN) techniques on these fields. Literature has shown that multi-head attention recurrent NN (MHA-RNN) in combination with late fusion (LF) [3, 9] approaches are the most prominent techniques. As discussed in [4] by N.Majumder et al., sequential (or hierarchical in their case) late fusion can filter out inter-modal correlation.

In most recent approaches [3, 4, 9], an off-the-shelf encoder is used for each modality. Then the encoding of each modality is being fused by an RNN. In [3], the authors use word2vec [5], OpenSmile [2] and 3D Convolutional NN (CNN) [10] for text, audio and visual modalities respectively, the encoding of which they use in a bidirectional MHA-RNN and a softmax layer at the end to perform binary sentiment classification. A similar approach is followed by Poria et al. in [6], but with the exception of using a custom CNN for the text encoding and a support vector machine (SVM) for the classification task.

In this paper, we propose a LF technique that uses the same uni-modal encoders as in [3]. Our technique is also based on an MHA-RNN model, however, the LF technique we use takes advantage of the sequential temporal dependencies across the modality encodings.

3 METHOD

3.1 Uni-modal Feature Extraction

In this work, we use LF combining uni-modal features extracted from 2199 short video utterances of the MOSI dataset [12]. The utterances were annotated with sentiment labels ranging from -3 (very negative sentiment) to +3 (very positive sentiment) and were produced by 89 speakers between 20-30 years old. From the video dataset, we extract the audio and then convert it into text using IBM Watson, a Speech to Text converter [1]. Each modality consists of sequential utterances that are fed into three different modality-specific encoders to extract the visual, audio and textual uni-modal features. For textual features we use the text-CNN model [5], for audio features we use OpenSmile [2] and for the visual ones 3D-CNN model [10]. All the above uni-modal extractors provide 300-dimensional features.

3.2 Fusion Technique

Our LF approach consists of two layers of LSTM nodes with multi-head attention (MHA-LSTM). The first layer (LSTM Block in Figure 1) contains three nodes, one for each of the three modalities, i.e. each node takes as input a sequence of utterances of a particular modality from textual, audio or visual features. We fuse the states of the first layer LSTM nodes in two steps:

1. First, we fuse the states of the LSTM nodes that correspond to the textual and audio features by calculating their inner product.

[1] https://speech-to-text-demo.ng.bluemix.net

PETRA 2021, June 29-July 2, 2021, Corfu, Greece

ACM ISBN 978-1-4503-8792-7/21/06…$15.00
https://doi.org/10.1145/3453892.3461009

2. Secondly, we calculate the inner product of the above outcome with the state of the LSTM node that corresponds to the visual features.

We observed that fusing textual and audio features first gives better performance. Consequently, the product of the second step is given as input to the second layer of our MHA-LSTM network (LSTM in Figure 1). Finally, the output of the MHA-LSTM network is given to a softmax layer which classifies among the 7 different review categories of the dataset.

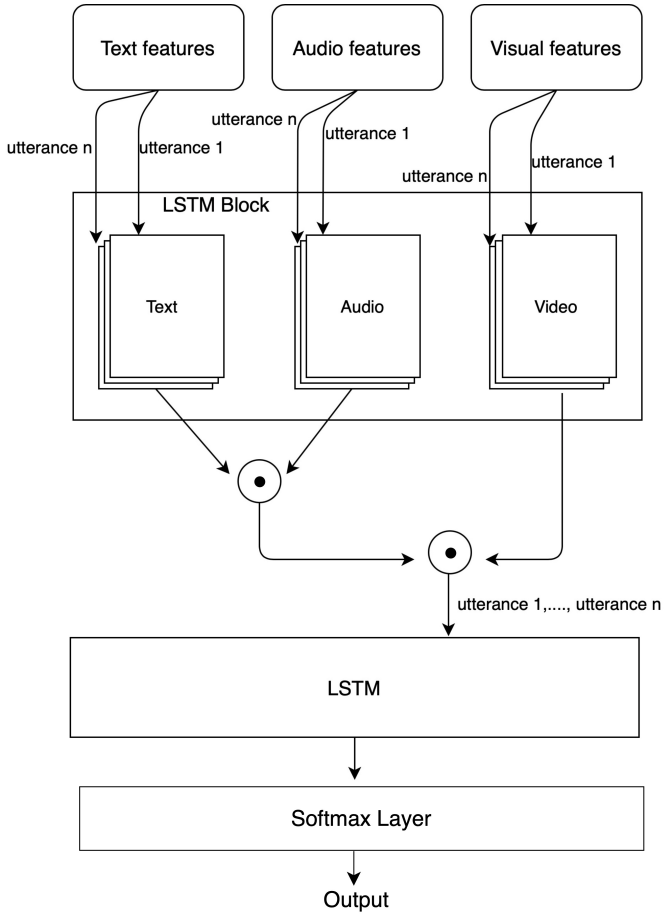

Figure 1: Fusion architecture using MHA-LSTM

3.3 Experiments

We train the MHA-LSTM network using the categorical cross-entropy as a loss function and we optimize using ADAM. Our model converges after 100 epochs. We split the dataset into 80% of training samples and 20% of testing ones. We evaluate our model in terms of accuracy, precision, recall and F1 score. Table 1 shows the detailed evaluation metrics.

Table 1: Performance metrics on our proposed model with MOSI dataset

Metric	Score
Accuracy	0.769
Precision	0.837
Recall	0.781
F1 Score	0.808

4 DISCUSSION AND FUTURE WORK

In this work, we propose a novel multi-modal SA method using LF with an MHA-LSTM model on uni-modal off-the-shelf feature extraction techniques. Our intention is to explore the potential of a different LF technique using a multi-head attention architecture. In the future, we plan to perform more extensive experiments with other datasets on SA exploiting again multiple modalities. These datasets will also involve other areas of sentiment like Post-Traumatic Stress Disorder and general stress.

REFERENCES

[1] Jorge A. Balazs and Juan D. Velásquez. 2016. Opinion Mining and Information Fusion: A survey. *Information Fusion* 27 (2016), 95–110. https://doi.org/10.1016/j.inffus.2015.06.002

[2] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia* (Firenze, Italy) *(MM '10)*. Association for Computing Machinery, New York, NY, USA, 1459–1462. https://doi.org/10.1145/1873951.1874246

[3] Taeyong Kim and Bowon Lee. 2020. Multi-Attention Multimodal Sentiment Analysis. In *Proceedings of the 2020 International Conference on Multimedia Retrieval* (Dublin, Ireland) *(ICMR '20)*. Association for Computing Machinery, New York, NY, USA, 436–441. https://doi.org/10.1145/3372278.3390698

[4] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems* 161 (2018), 124–133. https://doi.org/10.1016/j.knosys.2018.07.041

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs.CL]

[6] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain. 2018. Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines. *IEEE Intelligent Systems* 33, 6 (2018), 17–25. https://doi.org/10.1109/MIS.2018.2882362

[7] Arielle Scoglio, Erin Reilly, Jay Gorman, and Charles Drebing. 2019. Use of Social Robots in Mental Health and Well-Being Research: Systematic Review. *Journal of Medical Internet Research* 21 (07 2019), e13322. https://doi.org/10.2196/13322

[8] Jie Shen, Ognjen Rudovic, Shiyang Cheng, and Maja Pantic. 2015. Sentiment Apprehension in Human-Robot Interaction with NAO. https://doi.org/10.1109/ACII.2015.7344676

[9] Lukas Stappen, Alice Baird, Lea Schumann, and Björn W. Schuller. 2021. The Multimodal Sentiment Analysis in Car Reviews (MuSe-CaR) Dataset: Collection, Insights and Improvements. *CoRR* abs/2101.06053 (2021). arXiv:2101.06053 https://arxiv.org/abs/2101.06053

[10] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3D convolutional networks. In *2015 International Conference on Computer Vision, ICCV 2015 (Proceedings of the IEEE International Conference on Computer Vision)*. Institute of Electrical and Electronics Engineers Inc., 4489–4497. https://doi.org/10.1109/ICCV.2015.510 15th IEEE International Conference on Computer Vision, ICCV 2015 ; Conference date: 11-12-2015 Through 18-12-2015.

[11] A. Zadeh, R. Zellers, E. Pincus, and L. Morency. 2016. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intelligent Systems* 31, 6 (2016), 82–88. https://doi.org/10.1109/MIS.2016.94

[12] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. arXiv:1606.06259 [cs.CL]