MULTIMODAL ATTENTION NETWORK FOR TRAUMA ACTIVITY RECOGNITION FROM SPOKEN LANGUAGE AND ENVIRONMENTAL SOUND

By

RUIYU ZHANG

A thesis submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Master of Science

Graduate Program in Electrical and Computer Engineering

Written under the direction of

Ivan Marsic

And approved by

_____

_____

_____

New Brunswick, New Jersey

Jan 2020

ABSTRACT OF THE THESIS

Multimodal Attention Network for Trauma Activity Recognition from Spoken Language and Environmental Sound

by Ruiyu Zhang

Thesis Director: Prof. Ivan Marsic

Trauma activity recognition aims to detect, recognize, and predict the activities (or tasks) during a trauma resuscitation. Previous work has mainly focused on using various sensor data including image, RFID, and vital signals to generate the trauma event log. However, spoken language and environmental sound, which contain rich communication and contextual information necessary for trauma team cooperation, is still largely ignored. In this paper, we propose a multimodal attention network (MAN) that uses both verbal transcripts and environmental audio stream as input; the model extracts textual and acoustic features using a multi-level multi-head attention module, and forms a final shared representation for trauma activity classification. We evaluated the proposed architecture on 75 actual trauma resuscitation cases collected from a hospital. We achieved 72.4% accuracy with 0.705 F1 score, demonstrating that our proposed architecture is useful and efficient. These results also show that using spoken language and environmental audio indeed helps identify hard-to-recognize activities, compared to previous approaches. We also provide a detailed analysis of the performance and generalization of the proposed multimodal attention network.

ACKNOWLEDGEMENTS

We would like to thank the trauma experts from Trauma and Burn Surgery, Children's National Medical Center for their work on data collection and processing. This research was supported by the National Institutes of Health under Award Number R01LM011834.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Activity recognition in medical setting is challenging due to workflow complexity, fast pace, and environmental interference. Trauma resuscitation provides initial treatment of critically injured patients in an emergency, this particularly requires team dynamics and collaboration [1]. There has been much existing work using cameras, passive RFID, and medical equipment signals as input to detect and recognize clinical activity or phase [2–4]. but in this field, it is rare for human medical speech and environmental sounds to be used as input. Compared to other types of sensor data, speech and environmental sound contain extensive team cooperation information that indicates the performed tasks. For some specific activities such as *GCS calculation*, the trauma staff mainly relies on speech for communication. Ignoring this potentially important input source may complicate research in activity recognition in medical setting.

In this paper, we propose a deep learning neural network to recognize trauma resuscitation activities from verbal communication transcripts and environmental audio streams. Specifically, given a sentence-level verbal transcript and the corresponding audio stream from the trauma room, the proposed network outputs a trauma activity (shown in Fig. 1.1). There are two critical differences between our work and previous approaches: Firstly, instead of using cameras [3] and passive RFID [5,6], we use speech and environmental sound for activity prediction, to overcome the difficulty of recognizing speech-reliant activities. To the best of our knowledge, this is the first research that introduces an architecture using language information and context audio for trauma activity recognition. Secondly, other study [7] uses language to identify trauma phases, which are high-level states opposed to this paper's focusing on specific low-level activities. We have also taken environmental sound into consideration and therefore have built a multimodal model, which is more

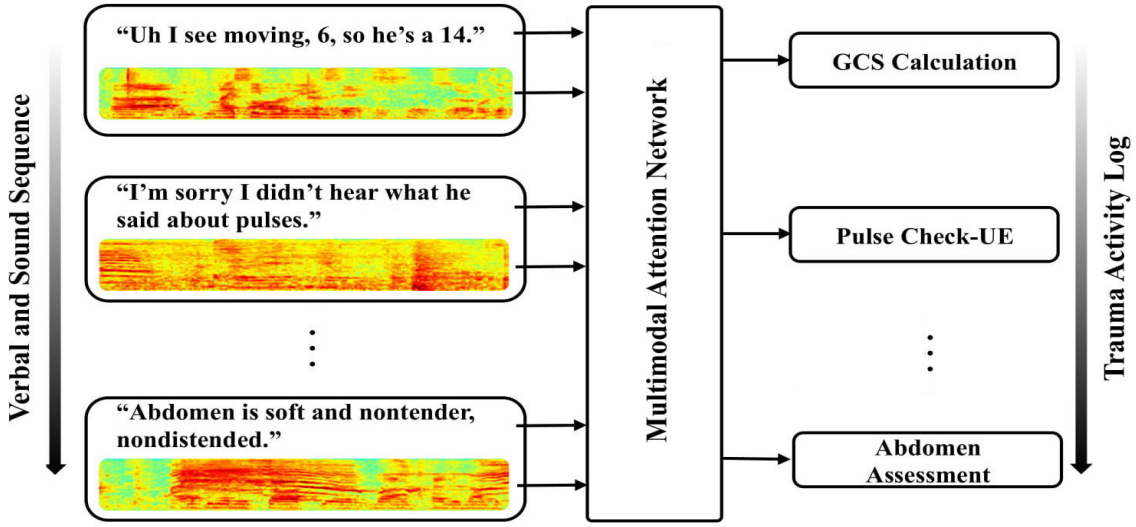

Figure 1.1: Example of spoken language and environmental sound based trauma activity recognition.

generalizable than a text-only model; the environmental sound can be seen as a complementary resource for the existing models. Our model accomplishes activity recognition in three steps: First, we process the audio stream and verbal transcript into spectrograms and text embeddings, respectively. Second, the model extracts feature representations from this preprocessed data using two multi-layer multi-head attention modules. Finally, we set up an attention-based fusion module to combine the modality-specific features, selecting representative and informative features. We directly connected the first and second step in the model and trained the system end-to-end.

We evaluate the proposed architecture on 75 actual trauma room resuscitation cases with recorded audio and spoken language transcripts. Both the audio stream and transcripts were segmented into sentence-level data; each sample contains one complete text sentence with the corresponding audio stream. Trauma experts assigned one of eleven different activity labels to each sample. We applied an 80%-20% training testing split with 5-cross

validation and considered the cases independently. The results show that the proposed multimodal attention network (MAN) achieves 71.8% accuracy with 0.702 F1-score in average, outperforming baselines with a more parameter-efficient model. The results also demonstrate the effectiveness of using speech and environmental sound as input sources for trauma activity recognition. Our contributions are:

- A multimodal architecture that considers spoken language and environmental sound to detect and recognize trauma resuscitation activities.
- An end-to-end multimodal attention network that automatically preprocesses raw data, extracts sentence-level acoustic and textual representations, fuses the feature vectors into a shared representation, and makes the final prediction.

The paper is organized as follows: chapter II describes the proposed structure in details. We discuss data collection and application in chapter III. We provide result analysis in chapter IV and limitation discussion in chapter V. We conclude in chapter VI.

CHAPTER 2
METHOD

The multimodal attention network (MAN) consists of three major modules: preprocessing, modality-specific feature extraction, and fusion (shown in Fig. 2.1).

2.1 Preprocessing

The input data includes both sentence-level verbal transcripts and audio stream. For verbal transcripts, as suggested in [8], we embed each word into a 200-dimensional *GloVe* vector [9], with unknown words randomly initialized. We allow embedding parameter tuning during the training stage, so that medical words sharing similar contexts will be located closely in the embedding space. All sentences are zero-padded with the max sentence length of 35.

We represent the audio stream as a spectrogram using Mel-frequency spectral coefficients (MFSCs). As demonstrated in [10, 11], MFSCs maintain the locality of the audio data and provide more detailed information compared to the Mel-frequency cepstrum coefficients. Following previous research [10], we use 40 filter banks to extract static from MFSCs. Instead of applying delta and double delta coefficients as in [11, 12], we only use the static coefficient set in chase of better performance of the static set under limited computation resource. Considering the maximum length of our MFSC feature maps is 600, we zero-pad and set up a hierarchical structure for the audio preprocessing. Unlike in [12], where attention weights are learned based on overall MFSCs, we believe the critical and relevant information in frame-level audio data only appear in the adjacent and nearby frames. It is difficult and inefficient to find dependencies between two distant audio frames; hence, we segment the MFSC feature maps into several 30-frame sub-maps. The final shape of each audio sample is (30, 40, 30), where the first index represents the number

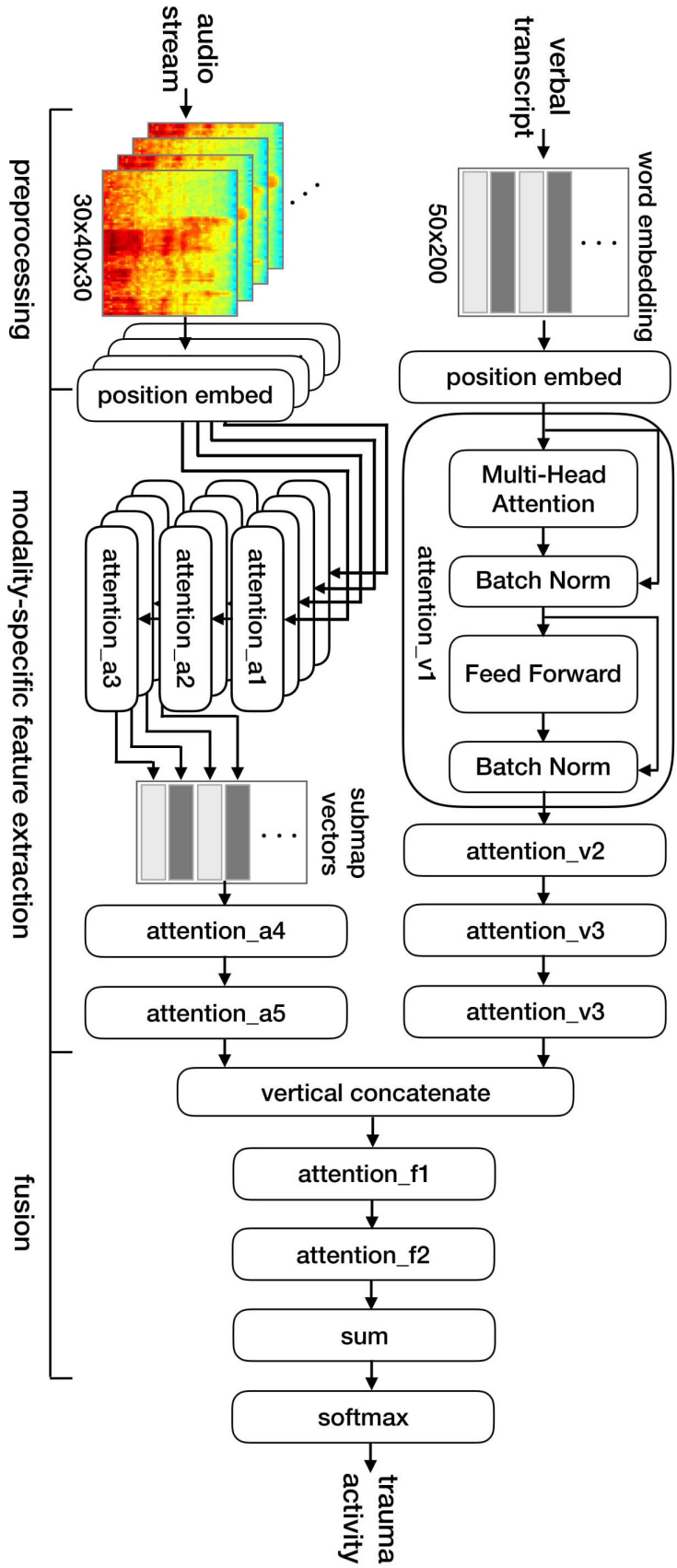

Figure 2.1: Overall structure of multimodal transformer network (MTN)

of the sub-maps, the second index indicates the energy frequency, and the last is the frame number of each sub-map.

2.2 Attention

Before bringing modality-specific feature extraction or fusion process into the topic, we briefly describe multi-head attention mechanism which is widely applied in our proposed model.

Attention was first introduced to help learn informative word representations in machine translation [13]. The function computes a weighted score to indicate the importance of each word, and the word representations weighted by their scores to form the final sentence representation. Multi-head attention [14] consists of several scaled dot-product attention layers in parallel to perform multiple attention computations for the input vector. Unlike general attention as in [15], multi-head attention applies scaled dot-product attention for each head based on the individual query, key, and value. It forms the final attention score by concatenating all the heads:

$$Q_i, K_i, V_i = xW_i^Q, xW_i^K, xW_i^V \tag{2.1}$$

$$Head_i(Q_i, K_i, V_i) = softmax(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i \tag{2.2}$$

$$y = Concat(Head_1, , Head_i, , Head_n)W \tag{2.3}$$

Where x is the input vector, and W_i^Q, W_i^K, W_i^V are the parameter matrices for the linear layer. The Q_i, K_i, V_i can be seen as the query, key, and value vector for the ith head. d_k is the dimension of the key. The final output is y. As mentioned in [14], the scaled dot-product attention is much faster and more space efficient. Compared to the general attention mechanism that learns the association based on the entire vector, the multi-head approach

improves the model performance by acquiring the information from various heads, each a sub-representation of the original vector.

Table 2.1: Modle parameters

*__input__=input shape; **output**=output shape; **n_h**=number of head; **h_s**=head size; **d_k**=dimension of key.

Layer	input	output	n_h	h_size	d_k
attention_v1	(50, 200)	(50, 160)	4	36	36
attention_v2	(50, 160)	(50, 100)	4	36	36
attention_v3	(50, 100)	(50, 60)	4	16	16
attention_v4	(50, 60)	(50, 30)	4	16	16
attention_a1	(30, 40, 30)	(30, 40, 30)	4	16	16
attention_a2	(30, 40, 30)	(30, 40, 30)	4	16	16
attention_a3	(30, 40, 30)	(30, 40, 30)	4	16	16
attention_a4	(30, 30)	(30, 30)	4	9	9
attention_a5	(30, 30)	(30, 30)	4	9	9
concatenate	(50|30, 30)	(80, 30)	-	-	-
attention_f1	(80, 30)	(80, 30)	4	9	9
attention_f2	(80, 30)	(80, 30)	4	9	9
sum	(80, 30)	(30)	-	-	-

2.3 Modality-specific Feature Extraction

The modality-specific feature extraction module has two independent networks to process the verbal transcript and audio stream, respectively.

Instead of using convolutional or recurrent neural networks (CNN/RNNs) [16, 17], we apply a multi-head attention network to extract the textual representations because: Firstly, sentence-level text classification requires focus on the most representative information, especially for short-sentence trauma speech. A single word can identify a specific class without using the rest of the text. For example, “GCS“ means *GCS Calculation* and “O_2“ means *Oxygen*. Replacing the CNNs and RNNs with attention concentrates on informative word vectors, rather than learning an entire sentence representation. Secondly, removing RNNs removes expensive in-sequence temporal alignment from the computation. The multi-head

attention model does not need the data fed in a specific order during the calculation. To provide temporal information, the model puts a position embedding layer before the attention function. In this research, we apply the same position embedding layer as in [14]. Considering the hardware performance tradeoff, we set four attention layers to extract representations from verbal transcripts. As suggested in [14], each attention layer consists of a multi-head attention module, a feedforward layer, and two batch normalization layers. Table 2.1 shows detailed model parameter information. It is worth mentioning that we designed a stepwise size reduction on the multi-head attention to improve model training and ensure matching dimensions between the transcript and audio feature representations.

As we mentioned in the preprocessing section, it is inefficient and unreasonable to compute dependencies across long-distance audio frames. Hence, we introduce a multi-level multi-head attention structure to first learn the attention distribution over adjacent audio frames, and then form the final feature vector over the entire MFSC map. We use three attention layers over each MFSC sub-map and further apply two extra attention layers to learn the consolidation of sub-map representations. The details of the parameters are shown in Table 2.1.

2.4 Fusion

The generated verbal and audio stream feature representations are of different length, so we concatenate them vertically to form the shared representation (shown in Table 2.1). We set two attention layers over the shared vectors to further fuse the features, which can be understood as weighing between verbal transcript and audio stream information together. The fusion attention layers select important features based on shared representations. We take the sum over the shared representations to form the final feature vector. A softmax classifier is used for the final classification.

CHAPTER 3
DATA COLLECTION AND IMPLEMENTATION

We collected 75 actual trauma resuscitation cases using two fixed NTG2 Phantom Powered Condenser shotgun microphones. Both microphones cover the major parts of the trauma room and have the ability to capture speech information and environmental sound from the trauma team. All the data were collected with consent, and have been stripped of private information manually by trauma experts (the medical team checked the data and manually muted the audio streams and removed the words that may involve or indicate privacy or personal information). We recorded the audio stream at a sampling rate of 16000Hz; the verbal transcripts were manually transcribed and segmented by the trauma experts; the activity labels were also provided by the medical team. The ten trauma activity labels are:

- *Back* (B)
- *GCS Calculation* (GCS)
- *Oxygen* (OX)
- *Head* (H)
- *C-Spine* (CS)
- *Pulse Check* (PC)
- *Blood Pressure* (BP)
- *Extremity* (E)
- *Mouth* (M)

- *Abdomen* (A)

All the rest utterances that do not belong to the above ten activities were assigned to *Other* (O) category. Table 3.1 provides detailed dataset statistics. We applied a 80%-20% training-testing split with 5-cross validation in experiment. For each training set, we further used 15% of the samples as validation set to help tune the model.

The model is implemented using *Keras* with *TensorFlow* as backend [18]. We first pre-train the audio branch for 50 epochs to facilitate model convergence, after which we trained the entire model for another 150 epochs. In order to overcome sample imbalance during training, we choose to uniformly sample across classes instead of directly feeding all the training data. For the entire training process, we also adopt dropout layer to help overcome model overfitting problem [19]. We first used Adam [20] optimization with 0.001 initial learning rate and momentum parameters 0.99 and 0.999 for the first 50 epochs. Then, we switched to SGD optimizer for further tuning.

Table 3.1: Dataset Statistics

Activity Type	**Number of Samples**
Extremity	731
Head	384
C-Spine	293
Blood Pressure	371
Back	582
Abdomen	265
Pulse Check	281
Oxygen	410
GCS Calculation	416
Mouth	282
Activity Labels in Total	4,015
Other	8,877
Labels in Total	12,892

CHAPTER 4
EXPERIMENT AND EVALUATION

We first made a quantitative analysis by comparing the performance of the modality-specific models and the multimodal structure. As is listed in Table 4.1, the verbal transcript model achieved 69.1% accuracy with 0.672 F1-score in average, while the environmental sound model only achieved 36.4% average accuracy with 0.342 average F1-score. Using verbal transcripts outperforms audio by 32.7% accuracy, this indicates that verbal communication from human speech contains information that is relatively more helpful to determine activity. This also shows that it could be difficult to identify trauma activity only based on environmental sound. Moreover, the multimodal structure outperforms the transcript-only model by 2.7% accuracy. On observing this, we believe the activity-specific medical machine sound or noise could possibly provide additional information to help improve overall performance of the model. The difference in performance demonstrates the necessity of multimodal architecture. Despite the limited performance of the audio-only model, the model based on combination of verbal information and environmental sound still shows the best performance.

Table 4.1: Comparison of modalities

Acc.=Accuracy (%).

Modality	***Data Type***	***Average-Acc.***	***Average-F1***
Verbal Transcript Only	Text	69.1	0.672
Audio Stream Only	Audio	36.4	0.342
Multi-modality (MAN)	**Text+Audio**	**71.8**	**0.702**

To further evaluate performance, we have studied the confusion matrices of our multi-modal attention network with the best performance training-testing split. As is shown in Fig. 4.1, *Blood Pressure* was classified most accurately, with an accuracy of 77.0%. Note

that the activities together classified as *Other* only can reach an accuracy of 55.0%, this result is lower than all the other classes. The explanation behind is that we only consider ten most common verbal-heavy activities and all the other activities are classified into the *Other* category, in this way we believe the internal diversity of the *Other* class makes it difficult to discriminate from the rest. However, the overall accuracies of the remaining activities are generally higher than 67.0%, this demonstrates the effectiveness of our MAN model.

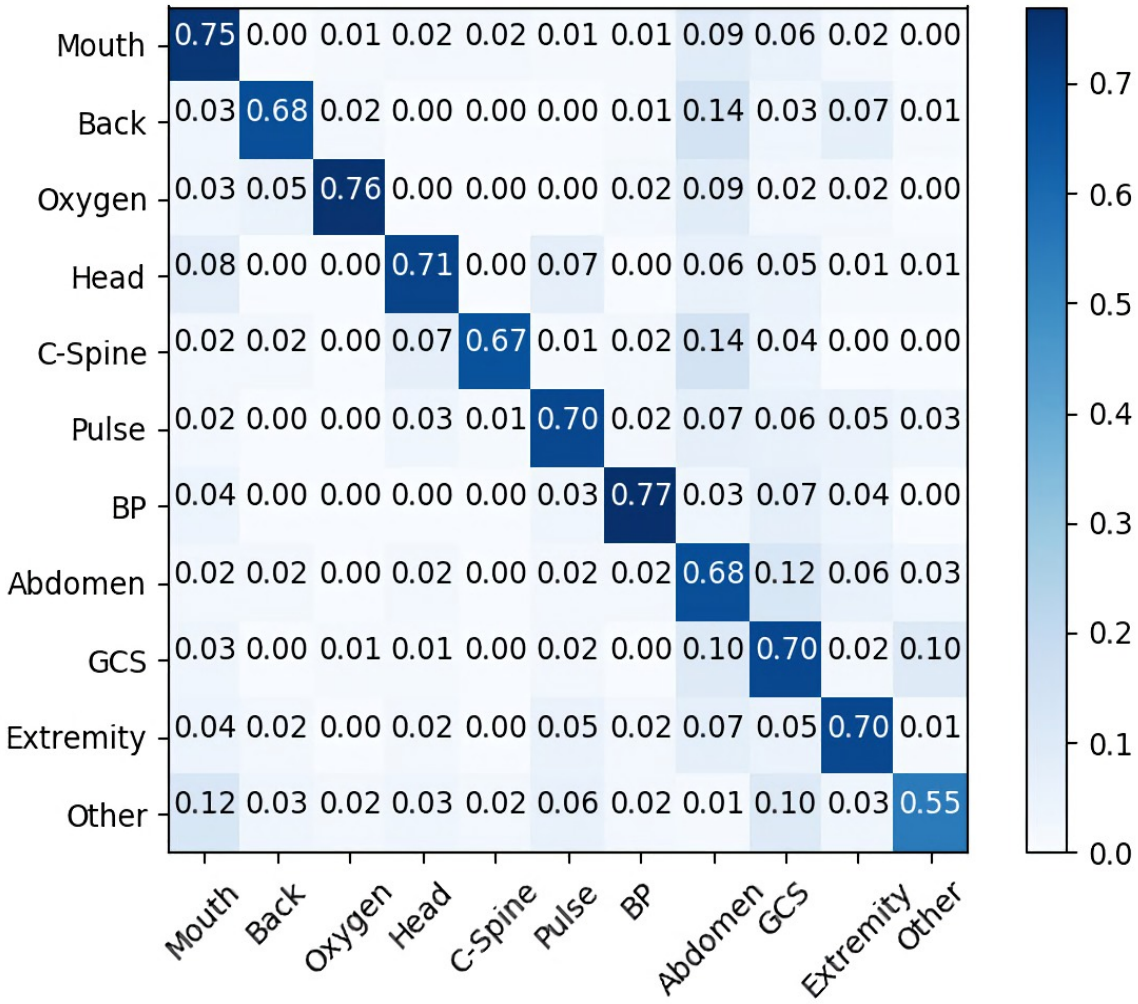

Figure 4.1: Confusion matrix of the MAN model.

To compare the proposed MAN with previous models, we first re-implemented the approaches in [7, 21]. Since the baseline approaches also used audio or text as input, we retrained them on our trauma dataset with exactly the same training-testing split. The result in Table 4.2 shows that the MAN model outperforms the baselines by 5.6% and 7.2% accuracy respectively. Because the distance between relevant sentences may vary in different cases, it is hard to define a fixed window size as in [7]. Compared to the hierarchical LSTM (H-LSTM) model that uses 20s as the context window size to predict the present activity, our model achieves better performance using only present verbal sentence without relying on any context information. Since text and audio data have less spatial features, using an attention network for feature extraction appears to be more reasonable than using convolution approaches. The result also indicates that our model significantly outperforms the H-CNN models proposed in [21], which demonstrates the effectiveness of MAN.

To better illustrate the necessity of using deep learning model, we also compared our model with several shallow and conventional models such as *SVM* and *Random Forest Tree*. We first concatenate the embedded textual word-level representations [9] and the low-level handcrafted acoustic features [22] as the joint features, and then we use the shallow *SVM* and *Random Forest Tree* classifiers to make the final decision. The result shows that our proposed MAN significantly outperforms these shallow models, this demonstrates that performing high-level feature extraction is much more effective than simply using low-level features of given data. Even with the limited data provided here, deep learning based models are still able to extract the representative features to improve the final classification results.

Considering the lack of RFID-based data in the experiment, we directly compared model performance on individual activities from [6] with our models in Table 4.3. The result shows that our model achieves better performance in three shared activities, including *Oxygen*, *Blood Pressure*, and *Mouth*. The MAN model gains a significant performance

Table 4.2: Comparison of baselines

Model	***Data Type***	***Accuracy (%)***	***F1-Score***
SVM	Text+Audio	55.4	0.512
Random Forest	Text+Audio	54.3	0.527
H-LSTM [7]	Text	66.2	0.623
M-CNN [21]	Text+Audio	64.6	0.642
Ours-MAN	**Text+Audio**	**71.8**	**0.702**

Table 4.3: Comparison of activities

Activity	**RFID in [6] (%)**	**Ours-MAN (%)**
Blood Pressure	64.1	**77.0**
Oxygen	54.0	**76.0**
Mouth	63.0	**68.0**
Pulse	**85.9**	70.0
Cardiac	92.9	-
Temperature	80.6	-
Ear	97.5	-
Warm Sheet	56.8	-
Nose	76.4	-
Pupils	59.6	-
GCS Calculation	-	70.0
Back	-	68.0
Head	-	71.0
C-Spine	-	67.0
Extremity	-	70.0
Abdome	-	68.0

improvement for the above activities, demonstrating the helpfulness of using verbal and environmental sound. Also, as is shown in Table 4.3, our model technically can not detect several types of activities such as *Ear*, *Nose*, *Pupils* etc. However, our model shows significant effectiveness on detecting activities like *GCS*, *Head*, and *Extremity*, which are difficult to detect using RFID. This result clearly indicates that spoken language and environmental sound can be applied as a complementary resource to improve trauma activity recognition as opposed to RFID-based data.

CHAPTER 5
LIMITATIONS

Even though the result demonstrates the effectiveness of MAN model, there still exists some limitations in terms of application. We listed three limitations below and treat them as focus of our future work.

First of all, this proposed MAN model is based on manually transcribed text as part of the input, which requires anticipation of human labor, or even trauma specialists, as we do involve such a team in our research. The automatic speech recognition (ASR) technology which allows speech-to-text without human transcripts could be a future option. Still, the performance of the ASR result relies heavily on the sound quality of the input audio stream. Considering that a trauma room or a surgery room could be noisy with various medical machine sound and irrelevant sobbing or crying sound from patients, together with "cocktail party problem", the ASR generated transcripts can hardly achieve an error rate as low as human transcription does. This strongly influences the performance of our model. Our future work will involve lowering the error rate of ASR transcription process. With that achieved, the time for data pre-processing could be lowered from days to minutes, and human labor cost could be eradicated.

Furthermore, the audio-only branch shows limited contribution compared to the verbal transcript branch does. Finding a more effective approach to extract the representative acoustic features from trauma resuscitation would still be an open-end topic. With a more detailed experiment design and analysis of the audio stream being achieved in the future, we will be able to deliver a breakthrough in terms of improving model accuracy and F-1 score.

Lastly, to design an applicable and scalable trauma activity recognition system, a combination of an effective RFID based model and this proposed MAN model is required. Our

future work should take into consideration a bigger picture which involves designing a generalizable multimodal system consisting of speech transcripts, audio stream, together with RFID signals. In this way, some certain types of activity that compatible better with RFID sensors will help improve the coverage of the whole system and a system with much higher overall effectiveness can be delivered.

CHAPTER 6
CONCLUSION

In this paper, we presented a novel approach using verbal communication information and environmental sound to recognize trauma resuscitation activities. We introduced a multimodal network with multi-head attention to extract and fuse textual and acoustic features. The proposed MAN achieved 72.4% accuracy with 0.705 F1 score. By outperforming the baselines, we demonstrate the effectiveness of the network and the necessity for the multimodal structure.

REFERENCES

[1] E.A. Bergs, F.L. Rutten, T. Tadros, P. Krijnen and I.B. Schipper, 2005. "Communication during trauma resuscitation: do we know what is happening?," Injury, 36(8), pp.905-911.

[2] J.E. Bardram, A. Doryab, R.M. Jensen, P.M. Lange, K.L. Nielsen and S.T. Petersen, 2011, March. "Phase recognition during surgical procedures using embedded and body-worn sensors," In 2011 IEEE international conference on pervasive computing and communications (PerCom) (pp. 45-53). IEEE.

[3] X. Li, Y. Zhang, M. Li, S. Chen, F.R. Austin, I. Marsic and R.S. Burd, 2016, October. "Online process phase detection using multimodal deep learning," In 2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON) (pp. 1-7). IEEE.

[4] N. Padoy, T. Blum, S.A. Ahmadi, H. Feussner, M.O. Berger and N. Navab, 2012. "Statistical modeling and recognition of surgical workflow," Medical image analysis, 16(3), pp.632-641.

[5] X. Li, D. Yao, X. Pan, J.Johannaman, J. Yang, R. Webman, A.Sarcevic, I. Marsic and R.S. Burd, 2016, May. "Activity recognition for medical teamwork based on passive RFID," In 2016 IEEE International Conference on RFID (RFID) (pp. 1-9). IEEE.

[6] X. Li, Y. Zhang, I. Marsic, A. Sarcevic and R.S. Burd, 2016, November. "Deep learning for rfid-based activity recognition," In Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM (pp. 164-175). ACM.

[7] Y. Gu, X. Li, S. Chen, H. Li, R.A. Farneth, I. Marsic and R.S. Burd, 2017, August. "Language-Based Process Phase Detection in the Trauma Resuscitation," In

2017 IEEE International Conference on Healthcare Informatics (ICHI) (pp. 239-247). IEEE.

[8] T. Mikolov, I.Sutskever, K. Chen, G.S. Corrado and J. Dean, 2013. “Distributed representations of words and phrases and their compositionality,“ In Advances in neural information processing systems (pp. 3111-3119).

[9] J. Pennington, R. Socher and C. Manning, 2014. “Glove: Global vectors for word representation,“ In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[10] O. Abdel-Hamid, A.R. Mohamed, H. Jiang, L. Deng, G. Penn and D. Yu, 2014. “Convolutional neural networks for speech recognition,“ IEEE/ACM Transactions on audio, speech, and language processing, 22(10), pp.1533-1545.

[11] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li and I. Marsic, 2018. “Multimodal affective analysis using hierarchical attention strategy with word-level alignment,“ arXiv preprint arXiv:1805.08660.

[12] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li and I. Marsic, 2018. “Hybrid Attention based Multimodal Network for Spoken Language Classification,“ In Proceedings of the 27th International Conference on Computational Linguistics (pp. 2379-2390).

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

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, . Kaiser and I. Polosukhin, 2017. “Attention is all you need,“ In Advances in Neural Information Processing Systems (pp. 5998-6008).

[15] Y. Gu, X. Li, K. Huang, S. Fu, K. Yang, S. Chen, M. Zhou and I. Marsic, 2018, October. “Human Conversation Analysis Using Attentive Multimodal Networks with

Hierarchical Encoder-Decoder,“ In 2018 ACM Multimedia Conference on Multimedia Conference (pp. 537-545). ACM.

[16] S. Lawrence, C.L. Giles,, A.C. Tsoi and A.D. Back, 1997. “Face recognition: A convolutional neural-network approach,“ IEEE transactions on neural networks, 8(1), pp.98-113.

[17] A. Graves, A.R. Mohamed and G. Hinton, 2013, May. “Speech recognition with deep recurrent neural networks,“ In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 6645-6649). IEEE.

[18] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard and M. Kudlur, 2016. “Tensorflow: A system for large-scale machine learning,“ In 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16) (pp. 265-283).

[19] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, 2014. “Dropout: a simple way to prevent neural networks from overfitting,“ The Journal of Machine Learning Research, 15(1), pp.1929-1958.

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

[21] Y. Gu, X. Li, S. Chen, J. Zhang and I. Marsic, 2017, May. “Speech intention classification with multimodal deep learning,“ In Canadian Conference on Artificial Intelligence (pp. 260-271). Springer, Cham.

[22] F. Eyben, M. Wllmer, and B. Schuller, 2010. ”Opensmile: the munich versatile and fast open-source audio feature extractor.” In Proceedings of the 18th ACM international conference on Multimedia (pp. 1459-1462). ACM.