M2D: A Multi-modal Framework for Automatic Medical Diagnosis

Raj Ratn Pranesh
Birla Institute of Technology
Mesra, India
raj.ratn18@gmail.com

Ambesh Shekher
Birla Institute of Technology
Mesra, India
ambesh.sinha@gmail.com

Sumit Kumar
Birla Institute of Technology
Mesra, India
sumit.atlancey@gmail.com

ABSTRACT

In this paper, we present M2D: a multimodal deep learning framework for automatic medical condition diagnosis via transfer learning. M2D leverages acoustic and textual features extracted from audio utterance and corresponding transcription describing a patient's medical symptoms. Our model utilizes ResNet-34 to learn audio feature via log mel-spectrogram and BioBERT language model to learn textual feature. We conducted a comparative performance analysis of M2D with baseline models based on textual or acoustic feature.

ACM Reference Format:
Raj Ratn Pranesh, Ambesh Shekher, and Sumit Kumar. 2020. M2D: A Multimodal Framework for Automatic Medical Diagnosis. In *Proceedings of Preprint*. ACM, 1 page. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Despite of recent advancements in healthcare and medical facilities, people all around the world are still facing medical resource scarcity and accessibility changes.

Recently various transfer learning based multimodal medical systems that uses combination of image, text and audio modalities has been developed. For making medical diagnosis convenient and accessible to a larger population, we designed a multimodal- M2D. Through M2D, we aim at providing automatic medical condition diagnosis system by analysing a given patient's symptoms recorded in the form of audio and text transcription. Our model utilizes a combined multimodal vector representation generated by the fusion of acoustic(log Mel-spectrogram) and textual modality vectors extracted via ResNet[1](a deep residual network) and BioBERT[3](a BERT model finetuned on large biomedical corpus) model respectively to classify a patient's symptoms into 25 possible medical condition classes.

2 METHODOLOGY

The M2D multimodal pipeline is consist of four stages. (i) **audio feature extraction**: From raw audio log mel-spectrograms is extracted(by *librosa* https://github.com/librosa) and supplied into ResNet model. The audio feature map(dimension = 512) was retrieved from the last Pooling layer of ResNet which is then passed into a dense layer which outputs a acoustic feature vector A_f with dimension of 256. (ii) **text feature extraction**: The audio transcript was tokenized(padded until max seq length = 30) and supplied into 12 attention-based encoding layers of BioBERT which finally outputs contextual word embeddings of size 768. This was then passed into a dense layer to produces a textual vector representation T_f of size 256. (iii) **multimodal feature fusion**: The acoustic and textual feature vectors were separately passed through a self-attention layer and then finally concatenated together to form a multimodal vector representation M_f of size 512. (iv) **classification**: M_f was passed through a dense layer(size=25) with softmax activation to make prediction.

3 EXPERIMENTS AND RESULTS

We used Appen dataset[1] having 6659 audio and it's transcription pair distributed over 25 medical diagnosis classes, divided in 70/20/10 ratio for train, validation and test. All of the dense layer had a RELU activation and a dropout layer(p = 0.5). Each model was trained for 30 epochs(batch size = 4) with learning rate of 1e-4 and Adam[2] optimizer. In our experiment 1, M2D outperformed both acoustic and textual based unimodel with an accuracy of **65.91%** which is 15.17% and 12.71% improvement over acoustic and textual unimodel respectively. Out of two unimodels, textual performed slightly better(2.81%) than acoustic model. In future we aim at extending our current work by experimenting with different combination of deep learning models and exploring other feature fusion techniques.

Model	Accuracy	Precision	Recall	F1
BioBERT	57.53%	58.03%	57.43%	57.73%
ResNet-34	55.91%	58.28%	55.21%	56.68%
M2D	**65.91%**	**68.91%**	**63.84%**	**66.23%**

Table 1: Models performance scores(in %)

REFERENCES

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[2] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[3] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[1] https://appen.com/datasets/audio-recording-and-transcription-for-medical-scenarios/

Preprint, 2021, Online

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn