EXPLAINABILITY OF DEEP MODELS
ADL PROJECT- II

Ritam Chattopadhyay, Kunal Sah

{ritamc, kunalsah}@iisc.ac.in

ABSTRACT

Deep learning models have been known as black boxes. In this project we try to explore two handpicked deep learning models and find explainability as to why the models behave the way they do. We work with video classification task using HMDB-51 Dataset and Language Identification task using Spoken Indian Language Identification Dataset. We extend the concepts of Grad-CAM, well known for image classification, on video classification task, and that of attention mechanism in finding relevance in language identification task.

***Index Terms*—** Video Classification, Language Identification, Attention, 2D Attention, Grad-CAM, LSTM, GRU, Explainability.

1. INTRODUCTION

Grad-CAM [1] by Selvaraju et. al. is technqiue to focus on the part of the image which the trained image classifier model has concentrated the most while classification. We extend the concept of Grad-CAM in the task of video classification. We use HMDB-51 (51 human action classes) video dataset to work on and CNN-LSTM based classification models in the lines of [2] Donahue et. al., for our task. For the language identification task we worked on Spoken Indian Language Identification Dataset from IEEE DataPort submitted by Sunil Kumar Kopparapu, following the methods LSTM, LSTM-Attention and LSTM-Attention2D.

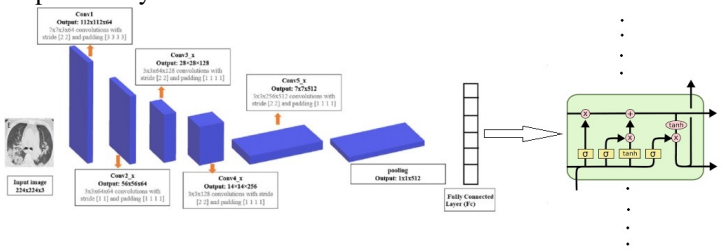
2. TECHNICAL DETAILS

2.1. Video Classification Task

Here are the two video classification models we work on , compare and find explainability of.

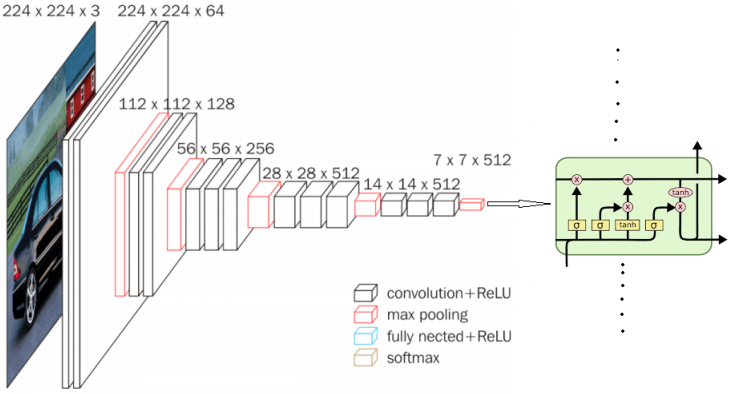
2.1.1. Using pretrained Resnet

In the CNN-LSTM based model [2] by Donahue et al, we use pretrained Resnet model to obtain the compressed version of each frame (spread out temporally) of a video and feed that vector to the LSTM [3]. The output vector spit out by the LSTM at the last time step is projected to 51 dimensions (denoting the number of classes). A softmax over this vector gives the probability distribution of different classes for that video.



2.1.2. Using pretrained VGG19

Similar to the above architecture, we use pretrained VGG19 in place of Resnet for model performance comparison. Rest of the architectural details remain the same.



2.2. Language Identification Task

For these task we have implemented three following models,

2.2.1. Using only BiLSTM

In the BiLSTM based model, MFCCs (Mel-frequency cepstral coefficients) of the audio files are used as the input to the BiLSTM. 128 mels is used in MFCC. Last 768 frames of the audio files are used to feed into the Model. In BiLSTM layer, 32 is used as the hidden dimension. After that using a Linear layer and a softmax layer, the probability distributions of the classes is given as the output.

Thanks to Google Colab platform for experiments

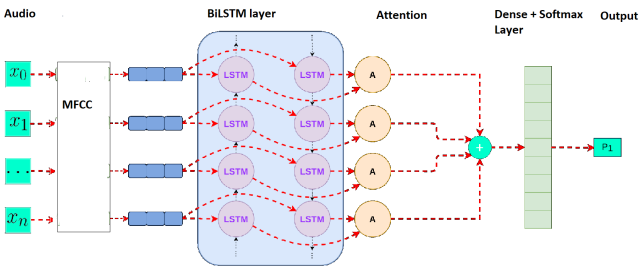


Fig. 1. BiLSTM-Attention

2.2.2. *Using BiLSTM-Attention*

Similar to the previous architecture, here after the BiLSTM layer an extra Attention Layer is used. Here attention of size number of frames(768)is used to determine which frame is to give more importance. After a Linear layer and a softmax layer is used to get the probability distributions of the classes.

2.2.3. *Using BiLSTM-Attention2D*

Similar to the previous architecture, here after the BiLSTM layer instead of one dimentional attention a two dimentional Attention Layer is used. The two dimentional attention is of size (number of frames[768])X(2*LSTM Hidden Dimention[64]) is used to determine which component of a frame is to give more importance.

3. CONTRIBUTIONS

- For the video classification task, we incorporate Grad-CAM for explainability which was predominantly introduced for images classification tasks.
- We introduce the 2D attention in the Language Identification task, which gives a clearer picture of explainability.

4. RESULTS

Video Classification Resnet-LSTM performed better on the test set than the Vgg-LSTM model. Accuracy of Resnet-LSTM model = 64%, while that of Vgg-LSTM model 32%.

Language Identification BiLSTM-Attention2D based model performed better than only BiLSTM and BiLSTM-Attention2D based models.

Model	BiLSTM	BiLSTM-Attention1D	BiLSTM-Attention2D
Accuracy	73.6%	86.1%	88.9%
CE Loss	0.505	0.257	0.107

5. RESOURCES

- Python, pandas, numpy, matplotlib
- Pytorch, sklearn, librospeech, OpenCv
- HMDB-51 dataset $< link >$
- Language dataset $< link >$

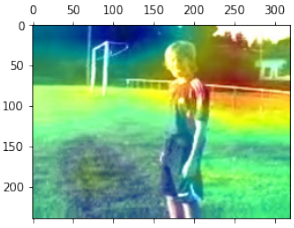
6. REFERENCES

[1] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

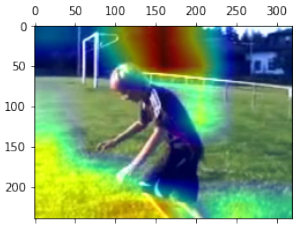
[2] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

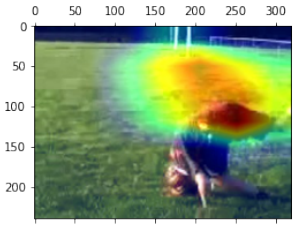
[3] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

7. APPENDIX

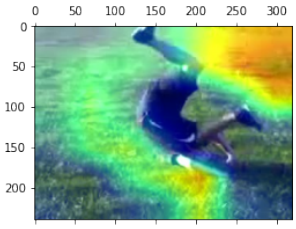


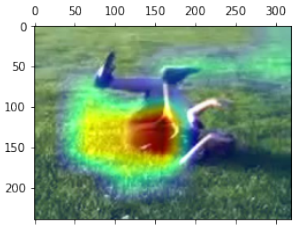
(a) Frame 1





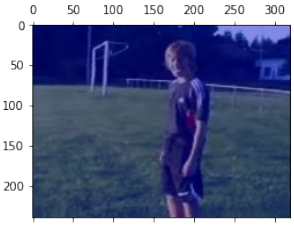
(b) Frame 2 (c) Frame 3



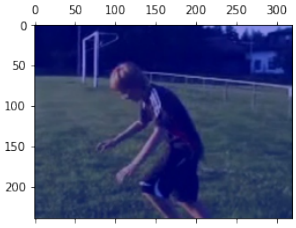


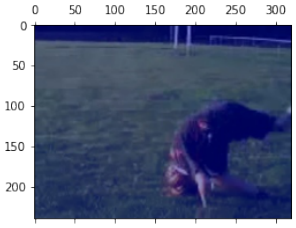
(b) Frame 4 (c) Frame 5

Fig. 2. Superimposed heatmap for somersault video classified by Resnet-Lstm model

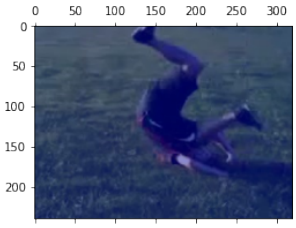


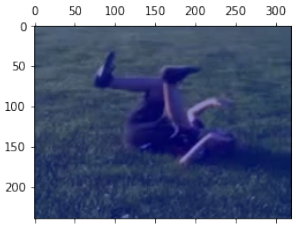
(a) Frame 1





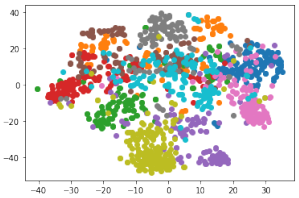
(b) Frame 2 (c) Frame 3



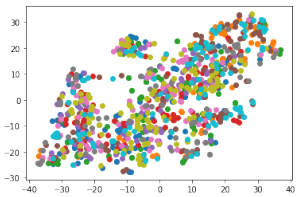


(b) Frame 4 (c) Frame 5

Fig. 3. Superimposed heatmap for somersault video classified by Vggnet-Lstm model



(a) Resnet-LSTM model



(b) Vgg-LSTM model

Fig. 4. t-SNE projection of the mean of the vectors (across time frames) spit out by LSTM for train set using different models

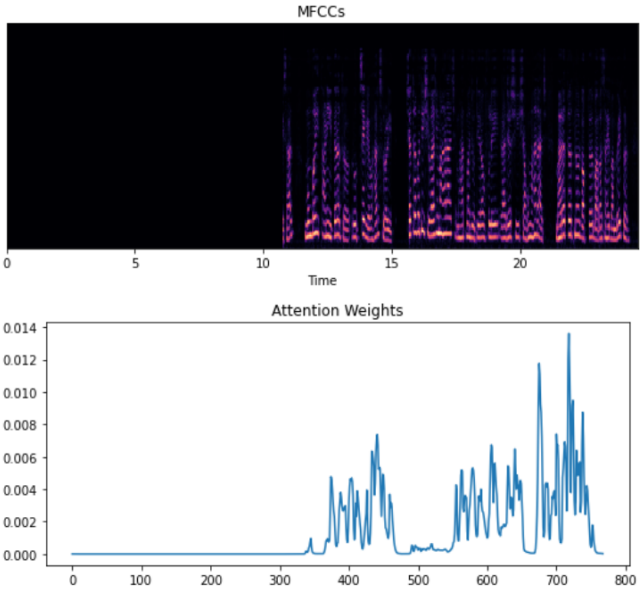


Fig. 5. Mfcc and corresponding attention weights

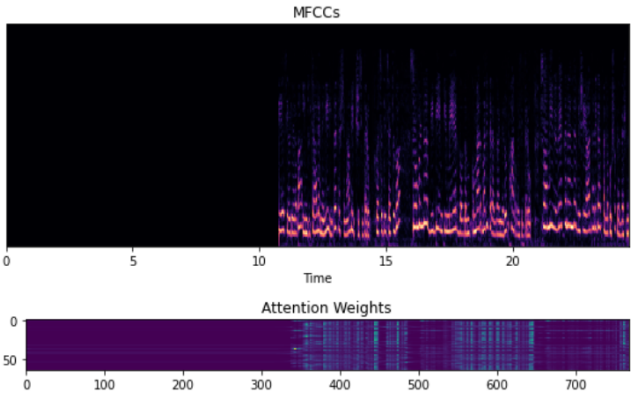
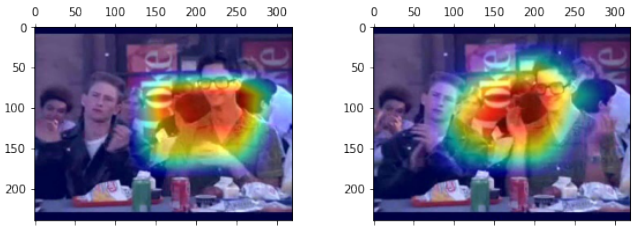
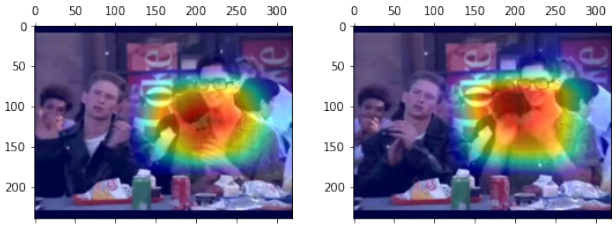


Fig. 6. Mfcc and corresponding two dimentional attention weights

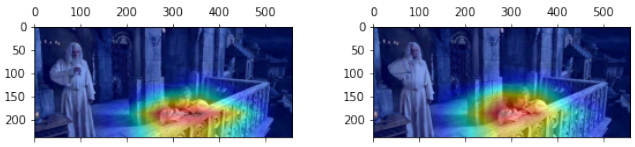


(a) Frame 1 (b) Frame 2

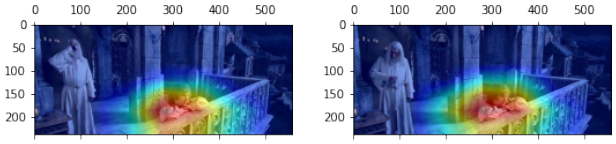


(c) Frame 3 (c) Frame 4

Fig. 7. An instance of correct classification
Correct Class: ”Clap”, Predicted Class: ”Clap”
(output by Resnet-LSTM model)



(b) Frame 1 (c) Frame 2



(c) Frame 3 (c) Frame 4

Fig. 8. An instance of incorrect classification.
Correct Class: ”Drink”, Predicted Class: ”Smile”
output by Resnt-LSTM model