Solving Multimodal Problem Using Deep Learning: Speaker Identification

Farzana Ahamed Bhuiyan, MD Bulbul Sharif

Tennessee TECH

Problem Statement

- Multimodal learning involves relating information from multiple sources
- Speaker Identification is a challenging multimodal problem integrating both visual and auditory signals
- Speaker identification refers to the task of locating the face of a person with the same identity as the voice in a video
- The process must be robust in dealing with severe degradations and unconstrained variations in content
- It needs to have a lower false alarm rate and higher recognition accuracy

Introduction

- Rapid progress in face recognition using Convolutional Neural Network (CNN)
- Rapid progress in speech recognition using Recurrent Neural Network (RNN)
- Speaker Identification, where both facial and auditory data is needed is much more challenging
- A multimodal LSTM architecture is the perfect choice to unify both visual and auditory modalities

Applications

- An important building blocks in many intelligent video processing systems such as video conferencing and video summarization
- By generate on-screen dynamic subtitles next to the respective speakers, it can enhance video accessibility for hearing impaired people
- Can enhance the overall viewing experience as well as reducing eyestrain for normal viewers.

Data & Feature Extraction

- Data from the "The Big Bang Theory" tv series
  - Six episodes from the first season is used as training data
  - Six episodes from the second season is used as test data
- Has different types of image degradation and high facial variations in the videos
- Feature Extraction for Videos:
  - Videos are nothing but a set of image collections
  - Extracted frames from the videos
  - Extracted all the faces of the characters from the frames using face detection algorithm
  - Manually gave labels to the images for training of the model
  - Converted each image into a fixed sized vectors using the InceptionV3 model
  - Get a 2048 length vector for each image

Fig 1: Extracted face image for the character "Sheldon"

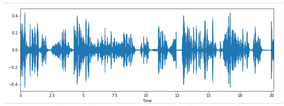


Fig 2: Extracted audio feature for the character "Sheldon"

- Feature Extraction for Audios:
  - Utilized the pre-annotated subtitles and only extracted audio segments corresponding to speeches
  - Loaded the audio data into a machine understandable format by taking values after every specific time steps (audio data sampling)
  - Used MFCC (Mel Frequency Cepstral Coefficients) coding to extract features from the audio representation
- Now both our image data and audio data are ready to fed into the neural network

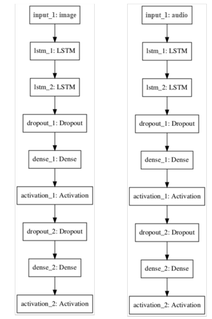
Methods

- Five alternative methods were used for the same task
- Single modal CNN method:
  - Trained a CNN for only face detection
  - Trained another CNN for only audio classification
  - Used a vote-based method to identify speaker
- Single modal LSTM method:
  - Trained a single modal LSTM for only face detection
  - Trained another separate single modal LSTM for only audio classification
  - Used a vote-based method to identify speaker
- Multimodal CNN-LSTM Method:
  - Used a single modal CNN for face images
  - Used a single modal LSTM for audio data
  - Merged the two modalities in the later stage, and created a multimodal CNN-LSTM classifier
- Early Fusion Multimodal LSTM Method:
  - Integrate data from the two domains and produce a larger input sequence
  - Used a multimodal LSTM for speaker identification using the combined data
- Late Fusion Multimodal LSTM Method:
  - Used two different single modal LSTM in parallel for two modalities
  - Merged the modalities in the later stage

Model Architecture

a. CNN Model for Face Detection b. CNN Model for Audio Classification

Fig 3: Single Modal CNN Models



a. LSTM Model for Face Detection b. LSTM Model for Audio Classification

Fig 4: Single Modal LSTM Models

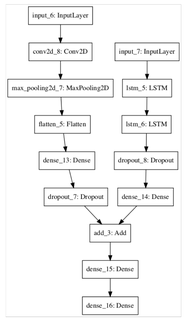


Fig 5: Multimodal CNN-LSTM Model

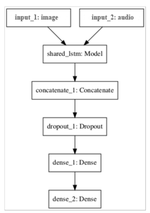


Fig 6: Early Fusion Multimodal LSTM Models

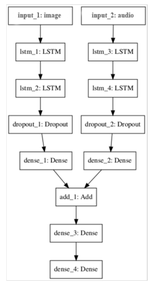


Fig 7: Late Fusion Multimodal LSTM Model

Results (Preliminary result)

Methods	Accuracy (%)	Precision	Recall	Area Under The ROC Curve
Single modal CNN method	73	78	61	63
Single modal LSTM method	65	69	53	57
Multimodal CNN-LSTM Method	88	86	78	71
Early Fusion Multimodal LSTM Method	91	93	87	88
Late Fusion Multimodal LSTM Method	86	86	79	68

Table 1: Comparison among the five methods with respect to accuracy, TPR, FPR and AUC

Discussion

- A big advantage of deep neural network approaches in data fusion is their capacity to learn from large amount of data.
- The major disadvantage of neural network approaches is their lack of interpretability. It is difficult to tell what the prediction relies on, and which modalities or features play an important role.
- Used both CNN and LSTM to work with speaker identification,
- Did a comparison among five different models, , and find out that the Late Fusion Multimodal LSTM model outperforms all other methods
- Proposed multimodal LSTM is robust again image degradation and distractors

Conclusions

- We proved that LSTM is the better choice to deal with sequence data
- We showed out multimodal LSTM did a good job in identifying speakers from video with high accuracy and low false positive rate
- We believe our multimodal LSTM is also useful to other applications, and not limited to the speaker identification task

Future Directions

- Comparison among No Cross-Modal Weight Sharing Method, Half Cross-Modal Weight Sharing Method and Full Cross-Modal weight Sharing Method
- Comparison with Multimodal CNN-LSTM method
- Overcoming problems such as noise robustness and variable channels
- Solving the task of speaker segmentation
- Indexing of multi-speaker speech

References

[1] J. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan, "Look, listen and learn multimodal lstm for speaker identification," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[2] Y. Hu, J. S. Ren, J. Dai, C. Yuan, L. Xu, and W. Wang, "Deep multi- modal speaker naming," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1107–1110.

[3] M. Everingham, J. Sivic, and A. Zisserman, "Hello! my name is... buffy"–automatic naming of characters in tv video." in *BMVC*, vol. 2, no. 4, 2006, p. 6.

[4] M. Tapaswi, M. Ba ̈uml, and R. Stiefelhagen, "knock! knock! who is it? probabilistic person identification in tv-series," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2658– 2665.

[5] A. Reynolds, "Automatic speaker recognition: Current approaches and future trends," *Speaker Verification: From Research to Reality*, vol. 5, pp. 14–15, 2001.

We would like to thank Dr. Doug Talbert for all of his help and support.