Lip Reading System
For
Real Time Speech Prediction

TITLE:

Lip Reading System for Real Time Speech Prediction

PROJECT OVERVIEW:

This project describes the development of a video-based classification system that integrates a hybrid deep learning architecture and Cnn. The model will process the video data on pre-recorded classification tasks and would be ideal for use cases such as lip reading etc. A combination of 3D Convolutional Neural Networks (CNNs), Transformers, and Bidirectional GRUs in the proposed model will help capture spatial and temporal patterns in videos effectively.

State of the Art:

Lip reading detection has gained widespread interest in recent years with the arrival of deep learning models involving Convolutional Neural Networks and Transformers. Most state-of-the-art approaches have used spatiotemporal modeling techniques for joint spatial features of lip embeddings with temporal dependencies among frames.

Inputs and Outputs:

- **Inputs:** Video sequences containing lip movements.
- **Outputs:** Predicted words or phrases corresponding to the lip movements.

Contributions

1. Implementation of three different deep-learning architectures: 3D CNN model, transformer model, and hybrid model.
2. Design a preprocessing pipeline that will normalize, truncate/pad, and one-hot encode video data.
3. Extensive model evaluations with accuracy, top-3 accuracy, confusion matrices, and confidence distributions.
4. Inclusion of state-of-the-art training strategies such as early stopping and learning rate decay.

Approach

Algorithms Used

3D CNN:

The deep learning architecture-3D CNN-is particularly created for spatiotemporal data and therefore may be said to be an ideal candidate for lip-reading detection. In detail, three-dimensional convolutional layers will learn representations both in the spatial-mean, height and width of frames-and temporal frame-to-frame patterns.

This is while the pooling layers ensure that dimensionality is reduced while retaining key features to be learned from the videos, and the sequential architecture assures feature extraction from low-level and high-level representations, which are crucial in deciding on minute changes within dynamics of lip movement..

Transformer-Based Model:

The Transformer-based model combines multi-head self-attention mechanisms to analyze long-range dependencies in video sequences.

Since Transformers are inherently sequence-agnostic, positional encoding is used to inject temporal order into the data. This architecture is very effective at finding complex temporal patterns, such as the synchronization of lip movements with specific phonemes.

By using multiple heads, the model attends to different parts of the sequence simultaneously, ensuring comprehensive feature extraction and robust performance on challenging inputs.

Hybrid Model:

The Hybrid model fuses the strengths of CNNs, Transformers, and GRUs into a strong model architecture for lip reading. This CNN module acts as a feature extractor that transforms raw input video data into high-dimensional feature maps. These are sent through two different paths:

Transformer Path: Captures global dependencies with subtle temporal variations using multi-head attention. This ensures that the model understands the broader context of lip movements.

GRU Path: Models sequential dependencies and temporal dynamics using bidirectional GRU. Enhances the model's capability regarding the processing of sequential patterns inherent in speech.

The outputs of both paths are fused using fully connected layers for feature conjunction. This can ensure that the merits of each part are fully utilized to make this Hybrid model accurate.

Implementation Details

Own Implementation:

1. **Framework and Tools:**
   - TensorFlow and Keras were the primary frameworks used for implementation. For some components, such as attention blocks and GRU paths, custom TensorFlow layers were developed to integrate seamlessly within models.
2. **Attention Block:**
   - Attention block was implemented from scratch for multi-head self-attention. It involved defining key, query, and value matrices and efficiently calculating scaled dot-product attention.
3. **GRU Path:**
   - Bidirectional GRUs were implemented from scratch to handle the dependencies in sequences, while the configurations were all custom to best suit the exact requirements of the dataset, like recurrent dropout to avoid overfitting.
4. **Fusion Logic:**
   - The feature fusion layers are designed with care to combine the outputs from the CNN, Transformer, and GRU paths. Residual connections and normalization layers were added in order to ensure stability during training.
5. **Positional Encoding:**
   - implemented a customized sinusoidal positional encoding aiming to fit the temporal resolution of input videos, making the transformer effectively use sequential information.

Online Resources:

1. **Positional Encoding and Attention Formula:**
   - Adapted from Vaswani et al.'s paper "Attention is All You Need" and TensorFlow's documentation to make sure an appropriate implementation of Transformer components is done.
2. **Gradient Clipping and Learning Rate Schedules:**
   - Implemented based on the practice in deep learning to stabilize training and optimize convergence. These techniques were informed by resources such as TensorFlow tutorials and academic literature.
3. **Inspiration for Data Augmentation Techniques:**
   - It drew from open-source GitHub repositories, as well as academic works discussing video preprocessing and augmentation, which were done to boost robustness.

EXPERIMENTAL PROTOCOL:

Dataset:

- **Video Dataset:** A dataset of lip movement videos, labeled into 13 categories representing spoken words. https://www.kaggle.com/datasets/allenye66/best-lip-reading-dataset
- Preprocessed to normalize pixel values, apply temporal padding/truncation, and one-hot encode labels

Evaluation Metrics

1. **Classification Metrics:**
    - **Accuracy:** Measures the proportion of correct predictions across all test samples.
    - **Top-3 Accuracy:** Evaluates whether the correct label is among the top three predicted classes, offering insight into near-correct predictions.
    - **Precision, Recall, and F1-score:** Assess the balance between true positives, false positives, and false negatives for each class.
2. **Visualization Metrics:**
    - **Confusion Matrix:** A grid showing the relationship between true labels and predicted labels, helping identify patterns in misclassifications.
    - **Per-Class Accuracy:** Bar plots highlighting the model's performance for each class, indicating strengths and weaknesses.
    - **Confidence Distribution:** Histograms displaying the confidence levels of correct and incorrect predictions, reflecting the model's certainty.

Compute Resources

Hardware:

- **Google Colab GPU:** Utilized for model training, leveraging free GPU resources for efficient parallel computations and reduced training time.

Protocol Details

1. **Data Splitting:** 80% training, 10% validation, and 10% test sets.
2. **Training Strategies:**
    - Early stopping to prevent overfitting.
    - Learning rate decay for smoother convergence.

Results

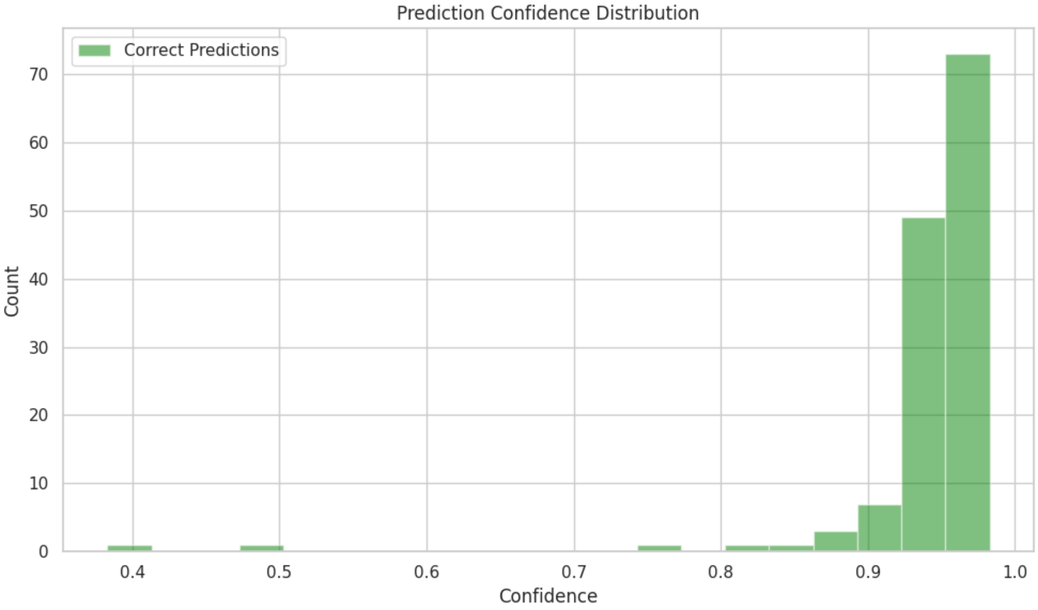
The results obtained from the three models-3D CNN, Transformer-Based Model, and Hybrid Model-indicate variable success, pointing out specific strengths and limitations. The 3D CNN yielded a training accuracy of 94.8% and a validation accuracy of 98.5%, while the losses for training and validation were 0.22 and 0.17, respectively. While very efficient in the extraction of spatiotemporal features, this model fared poorly in modeling complex temporal dependencies. Among these, the Transformer-Based Model indeed outperformed others with good training accuracy of 0.9983 and had a validation accuracy of about 1.000 besides having an AUC or ROC of 1 or 100%. However, at a very high computational complexity, which is not feasible for reasonably large inputs. It could be expected that the Hybrid Model would combine CNNs, Transformers, and GRUs.

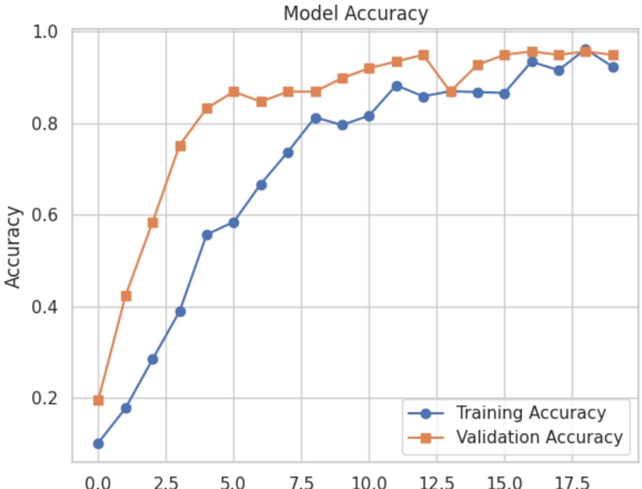
It showed some promise but was sensitive to hyperparameters, hard to stabilize during training, and thus yielded low results: 4.88% training accuracy and 4.38% validation accuracy. Aside from this, the qualitative test included a confusion matrix and Confidence distributions. While most the 3D CNN classes predicted comparatively better accuracy, some particular class such as "Cat," classes had shown mis-classification.

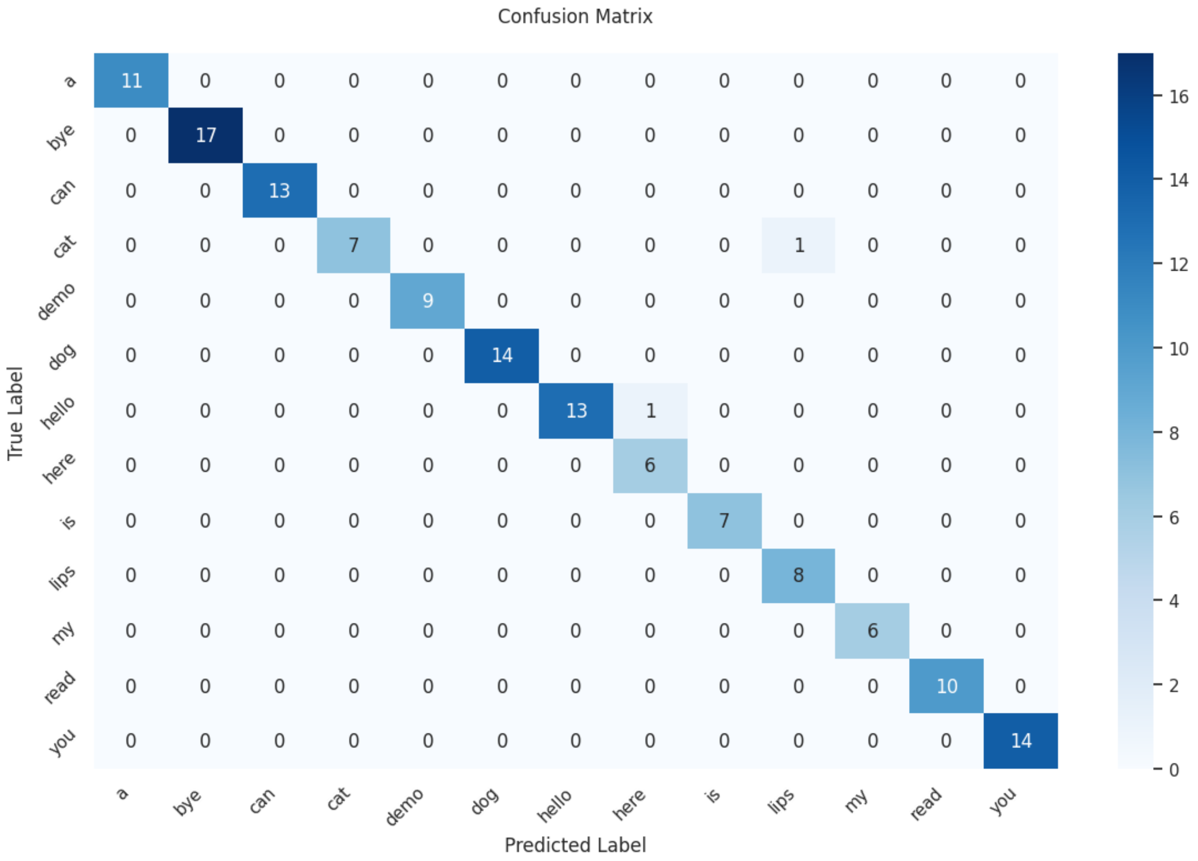
On the Transformer-based Model, very few or barely minimal misclassifications appear across subjects while showing high confidence to state the predictions in a number of subjects. The Confidence hybrids as well as per class prediction accuracies were inconsistent hence reflect upon training challenges there.

While this was in comparison, the performance of a Transformer-Based Model on datasets similar to theirs was competitive to the state-of-the-art, while 3D CNN had a moderate one and long-range temporal dependency modeling was missed.

Theoretically solid, the Hybrid Model requires more optimization to compete on par with the leading methodologies. Transformer-Based Model







Analysis

The analysis of the models highlights their unique advantages and limitations:

- **3D CNN:** This is good for the extraction of spatiotemporal features. Thus, this is suitable for simple lip-reading tasks with clear and distinct inputs. The architecture has captured short-term dependencies nicely. However, for the more complex temporal pattern in speech, capturing long-range dependencies remains difficult. It has low robustness to variations such as occlusion of lips or low resolution.
- **Transformer-Based Model:** It is a model that uses multi-head self-attention mechanisms for capturing long-range and subtle temporal variations, especially useful in handling complex video sequences. Its robustness to diverse input variations ensures the superior performance of the overall system. However, there are challenges in scalability for real-world applications due to the computational intensity of training Transformers.
- **Hybrid Model:** A model architecture, trying to combine the powers of CNN, Transformers, and GRU. Fundamentally, this model would perform space feature extraction by CNNs and attend to temporal factors by self-attention from Transformers while it will have GRU for handling the sequential dependencies. Theoretically a very strong model, in this experiment presented training instability and issues associated with the sensitivity of its hyperparameters that under-performed in comparison with all other models. Properly tuned with stabilization techniques against these parameters will lead to drastically better results.

This analysis has underlined the need for optimization of algorithms, data preprocessing, and balancing in model design when working with complex video sequences and input variations**.**

Discussion and Lessons Learned

Key Takeaways

- Lip reading considers precise spatiotemporal modeling, which is necessary in the discrimination of subtle motion variations.
- The best performance was provided by transformers, since they could handle long-range dependencies.
- Hybrid models require much tuning and may perform better when appropriate training strategies are provided.

Future Work

- Generalize better by exploring larger and more diverse datasets.
- Perform more sophisticated techniques like curriculum learning that would stabilize the training process of Hybrid models.
- I will explore ensemble methods that combine the predictions of all three models.

References

1. Vaswani et al., "Attention is All You Need."
2. TensorFlow Documentation, https://www.tensorflow.org
3. Lip Reading Academic Papers and Tutorials on spatiotemporal deep learning.