Title: Lip reading using computer vision

Name:

- Sachin Shivanna – A-20552795
- Harshith Deshalli Ravi – A-20552830

Main paper name: LipNet: End-to-End Sentence-level Lipreading

Publication Details:

- *Authors***:** Yannis Assael, Brendan Shillingford, Shimon Whiteson, Nando de Freitas
- *Journal***:** arXiv
- *Submission Date:* November 5, 2016 (v1), with a revision on December 16, 2016 (v2)
- *Abstract:* The paper presents, the first end-to-end sentence-level lipreading model that learns spatiotemporal visual features and a sequence model simultaneously. It achieves a 95.2% accuracy on the GRID corpus for sentence-level tasks.

Problem Statement:

The rise of digital editing tools has made it easy to manipulate videos by altering their audio content. A common deceptive technique involves taking a genuine video and replacing its original audio with different speech or sounds, creating misleading content that can broadcast misinformation. This presents a growing challenge for viewers trying to determine whether the audio and video components of media are genuine and properly synchronized.

To combat this issue, there's a need to develop automated systems that can verify whether a video's lip movements truly match its audio track. Such tools would be valuable for detecting manipulated content and helping maintain the trustworthiness of digital media.

Current approaches to this problem have limitations. Many existing systems focus on identifying individual words rather than understanding complete sentences in context. This narrow word-by-word approach makes it harder to accurately assess whether speech matches lip movements over time, since natural speech requires understanding the flow and timing of entire sentences.

Approach:

Our research proposes a novel deepfake detection system that analyzes the correlation between lip movements and speech in video content. We'll build upon existing visual speech recognition technology, specifically adapting the LipNet architecture, to identify inconsistencies between visual and auditory components.

The implementation strategy consists of four key phases:

1. Dataset Construction and Preprocessing
   - Build a comprehensive video collection featuring natural speech patterns
   - Generate controlled test cases by creating synthetic misalignments
   - Process the footage to standardize input formats and extract relevant features
2. Model Development and Optimization
   - Modify the core LipNet architecture to focus on synchronization analysis
   - Incorporate both spatial and temporal convolution layers
   - Implement sequence modeling using bidirectional recurrent networks
   - Design custom loss functions to emphasize temporal alignment
3. Discrepancy Analysis Framework
   - Develop algorithms to quantify the alignment between predicted speech patterns and actual audio

   - Create a scoring system for deviation severity
   - Set dynamic thresholds for flagging suspicious content
   - Build real-time processing capabilities
4. Validation and Testing Protocol
   - Design comprehensive test scenarios covering various speaking conditions
   - Measure detection accuracy using industry-standard metrics
   - Conduct ablation studies to optimize model components
   - Assess performance across different types of manipulated content

This system aims to strengthen content authenticity verification by introducing robust methods for detecting audio-visual inconsistencies in digital media. The framework will provide content moderators and verification systems with an additional layer of automated screening capability.

Supplementary Paper:

- *Paper:* [*https://ieeexplore.ieee.org/document/8373881*](https://ieeexplore.ieee.org/document/8373881)
- *Paper Name:* LCANet: End-to-End Lipreading with Cascaded Attention-CTC
- *Publiction:* IEEE
- *Author:* Kai Xu , Dawei Li , Nick Cassimatis , Xiaolong Wang

Improve technique:

To enhance the LipNet model, we propose integrating techniques from the LCANet paper, where we are focusing on cascaded attention on Connectionist Temporal Classification (CTC) optimization. We believe using this architecture will significantly improve the decoder decision and more robust for continuous speech recognition.

Data:

- *Training Data :* mpeg file with audio and video data.
- *Test Data:* mpeg file with video data.

Inspiration:

- *References paper:*
  1. Wand et al., 2016
  2. Chung & Zisserman, 2016a
  3. Eston & Basal, 1982

Gergen et al., 2016

- *Web sources:*
  5. arXiv page for the LipNet paper: https://arxiv.org/abs/1611.01599

- *Software:*
  6. TensorFlow (for model fitting and monitoring)
  7. Open CV
  8. Numpy
  9. Matplotlib

- *Data sources:*
  10. GRID corpus - used for training and evaluating LipNet
  11. MIRACL dataset - mentioned as an alternative dataset used for small model preparation

- *Additional resources:*
  12. Face detection and alignment scripts provided in the GitHub repository
  13. Pretrained weights for the LipNet model

Team Responsibility:

- *Sachin Shivanna:* Data cleaning and pre-processing. Creating data pipeline.
- *Harshith Deshalli:* Model training and testing, performance evaluation.