PURDUE
UNIVERSITY

Lip-to-Text for the Hearing Impaired: Multi-Modal Approach Using Vision-and-Language Transformer

CNIT 581-AST Project, Fall 2022

Nadine, Srushti, Yi

December 6, 2022

Agenda

- Introduction
  - Motivation
  - Existing Solutions
  - Targeted Gap
  - Objective
- Backgrounds
  - CNN
  - LSTM
  - ViLT
- Approaches
  - Dataset
  - Data Preprocessing #1
  - Data Augmentation
  - Data Preprocessing #2
  - Experiment Setup
- Results & Analysis
  - Results Comparison
  - Zero-Shot Results
  - Fine-Tuning Results
  - Demo Video
- Conclusions & Future Works

Introduction

Motivation

Lip-to-Text for the Hearing Impaired: Multi-Modal Approach Using Vision-and-Language Transformer

1.5 billion people with hearing loss,
25% of people over 60 years
(WHO, 2021)

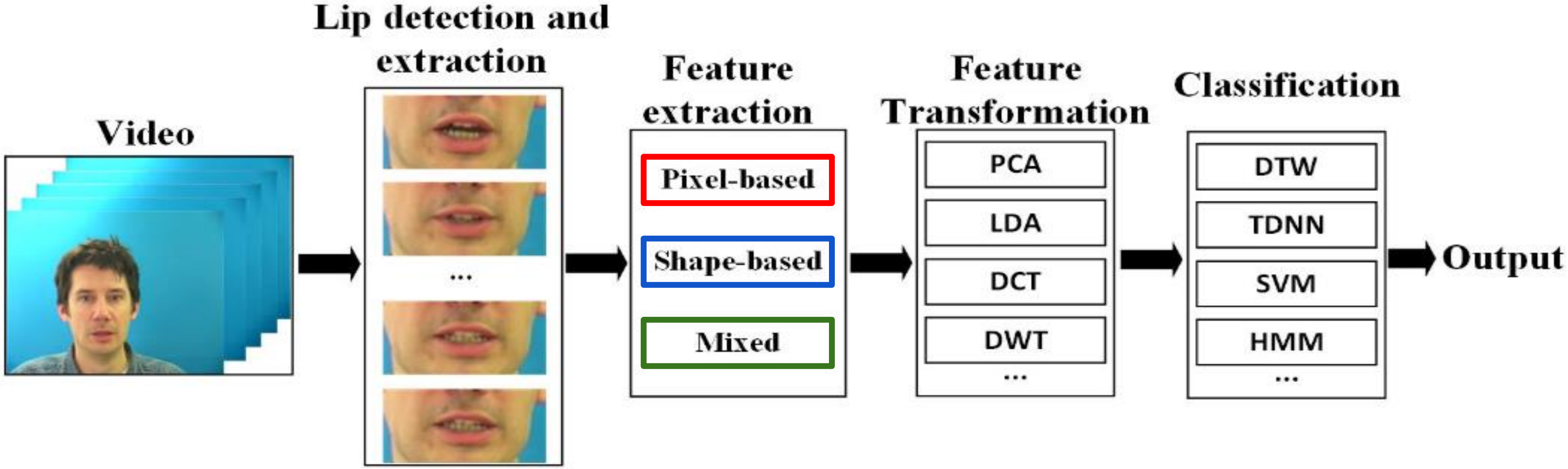
heavy dependence on lip reading & multi-tasking can be impractical

considering several modalities boosts performance

efficiency and speed are crucial

Existing Solutions

- ## Traditional Lip Reading



Morade & Patnaik, 2014
Sterpu & Harte, 2017

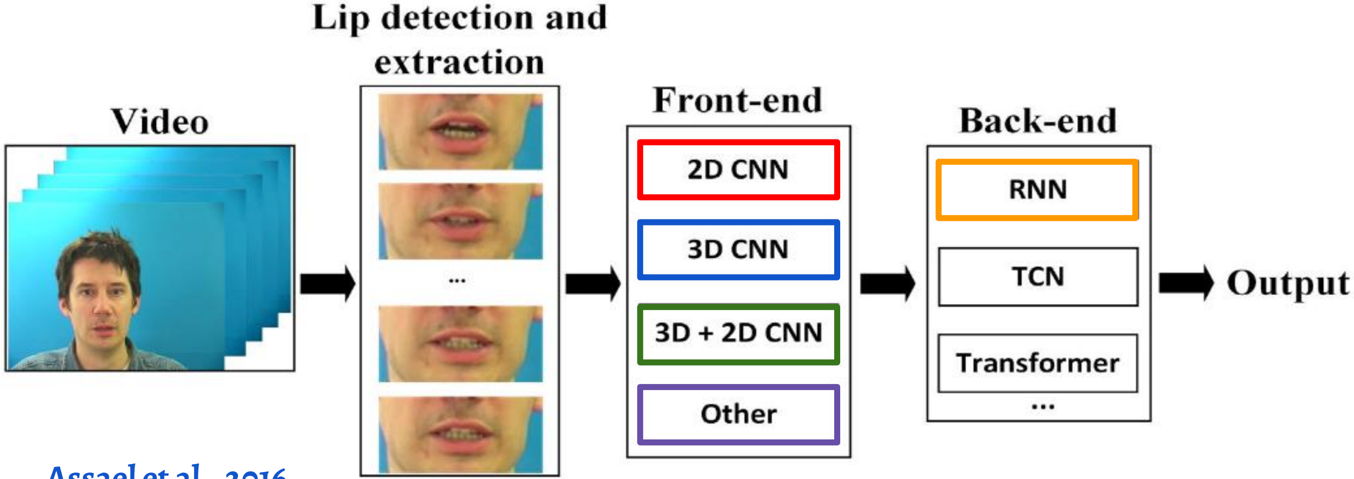
Luettin and N. A. Thacker, 1997
Ma et al., 2016

Bear et al., 2017
Howell et al., 2016
Watanabe et al., 2016

(Hao et al., 2020)

Existing Solutions

- ## Deep Lip Reading



Garg et al., 2016
Li et al., 2016
Mesbah et al., 2019
Noda et al., 2014
Saitoh et al., 2016
Zhang et al., 2019

Assael et al., 2016
Fung & Mak, 2018
Qiu et al., 2017
Torfi et al., 2017
Tran et al., 2017
Yang et al., 2019

Margam et al., 2019
Petridis et al., 2018
Stafylakis & Tzimiropoulos, 2017

FNN: Wand et al., 2016, 2017 & 2018
Autoencoder: Petridis et al., 2017 & 2018

Bi-LSTM: Stafylakisa et al., 2018; Weng & Kitani, 2019
Bi-GRU: Luo et al., 2020; Xiao et al., 2020;
Zhao et al., 2020; Zhang et al., 2020

(Hao et al., 2020)

Targeted Gap

The systems heavily rely on computationally complex feature extraction from visual input.

affects efficiency & speed of overall system

(Kim et al., 2021)

Objective

Evaluate the performance of the Vision-and-Language Transformer (ViLT) model, which provides a shallow, convolution-free embedding of input pixels, in the lip-reading task.

Our focus is on fine-tuning and testing the ViLT model on a publicly available lip-reading dataset; we are not concerned with how the input data is obtained or output is displayed in real-time.

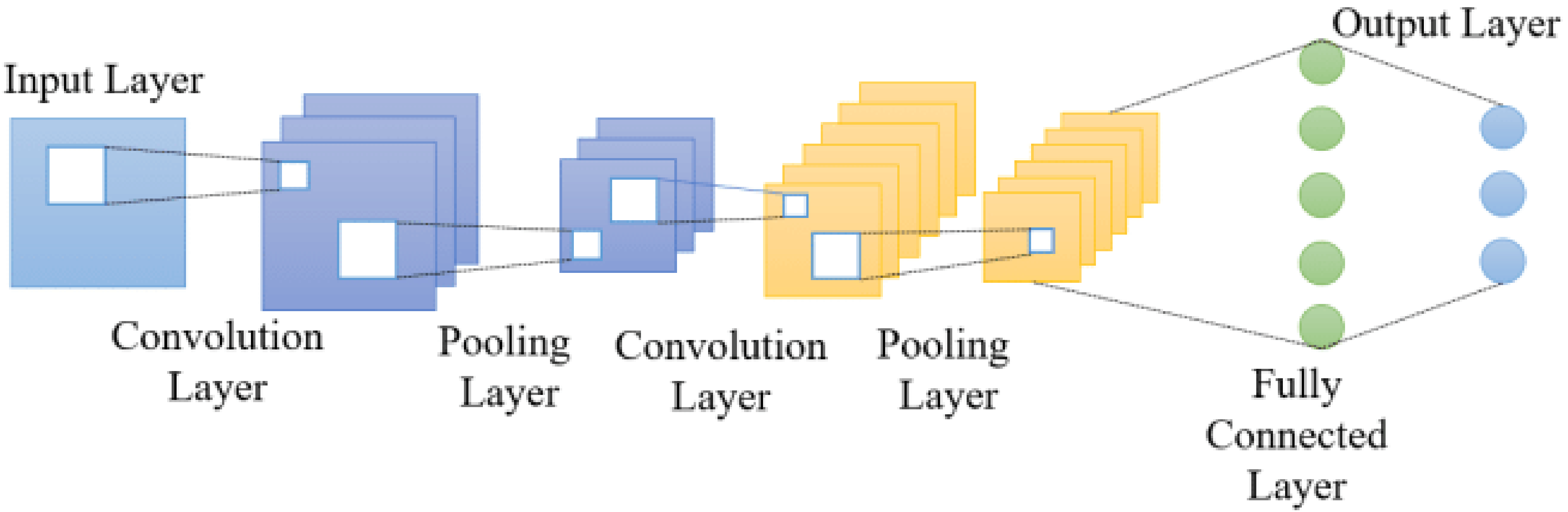
Backgrounds

Backgrounds

- Convolutional Neural Network (CNN):
    - Ideal for computer vision, classification, and object detection tasks
    - Several Layers of interconnected nodes
    - Results are based on extracted features

Backgrounds

- **Convolutional Neural Network (CNN):**



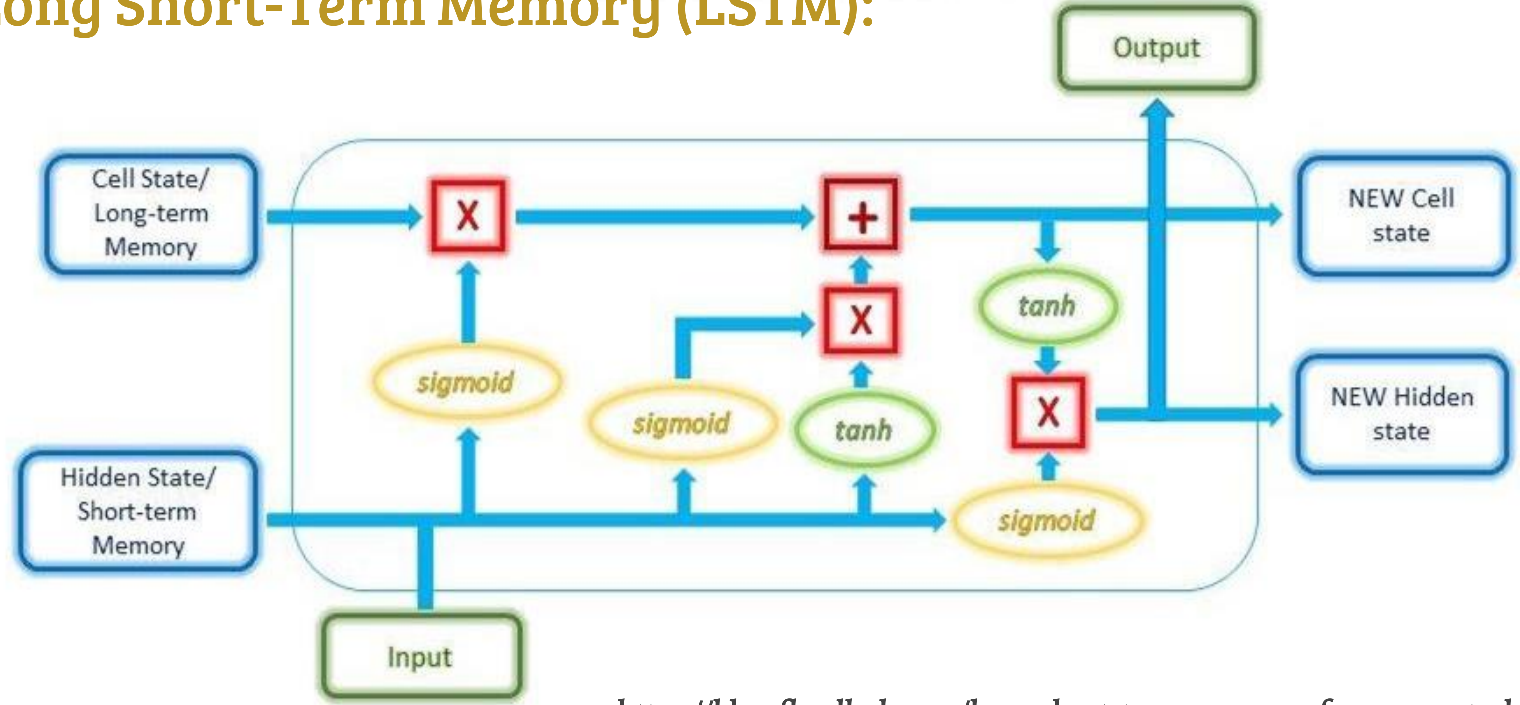
(Gu et al., 2019)

Backgrounds

- **Long Short-Term Memory (LSTM):**
  - Ideal for processing sequential data
  - Chain structure
  - Commonly used for language translation and text generation

Backgrounds

- **Long Short-Term Memory (LSTM):**

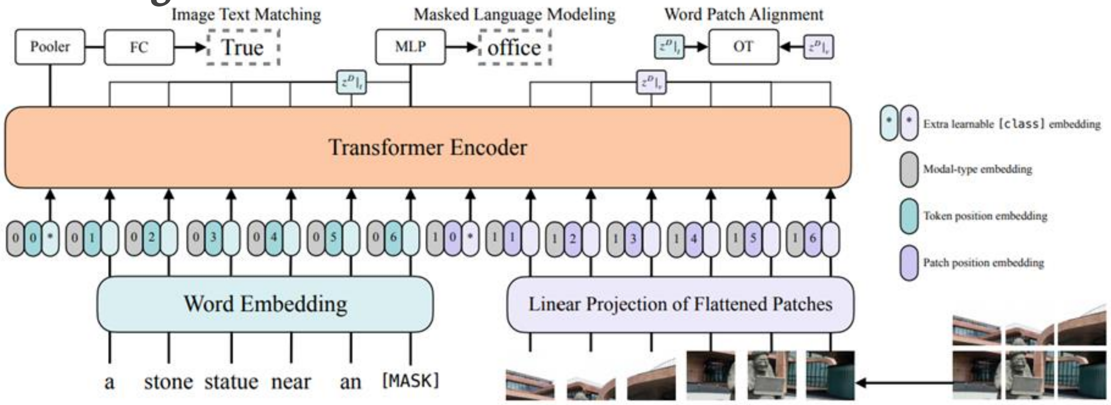


https://blog.floydhub.com/long-short-term-memory-from-zero-to-hero-with-pytorch/

Backgrounds

- **Vision and Language Transformer (ViLT)** (Kim et al., 2021)
  - Trained on 4 datasets: COCO, Visual Genome, Conceptual Captions & SBU Captions
  - 2 pre-training tasks:
    - Image Text Matching
    - Masked Language Modeling

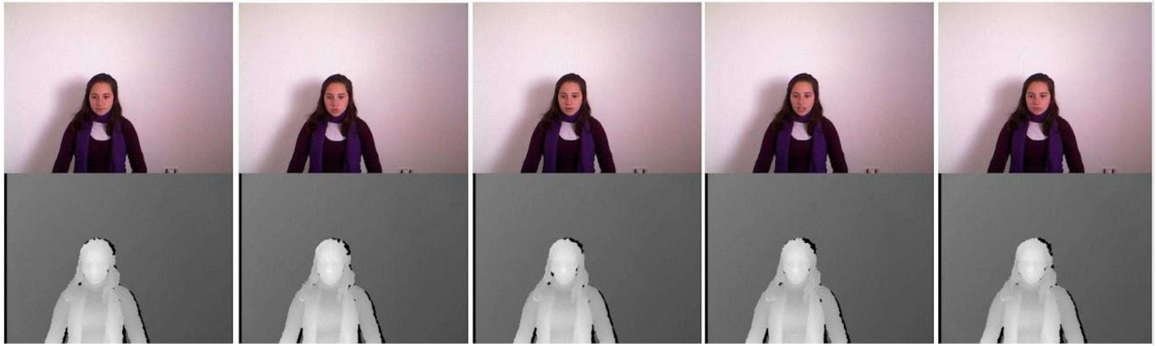
Visual Embed	Model	#Params (M)	#FLOPS (G)	Time (ms)
Region	ViLBERT	274.3	958.1	~900
	UNITER	154.7	949.9	~900
Linear	ViLT	87.4	55.9	~15



Approaches

Dataset

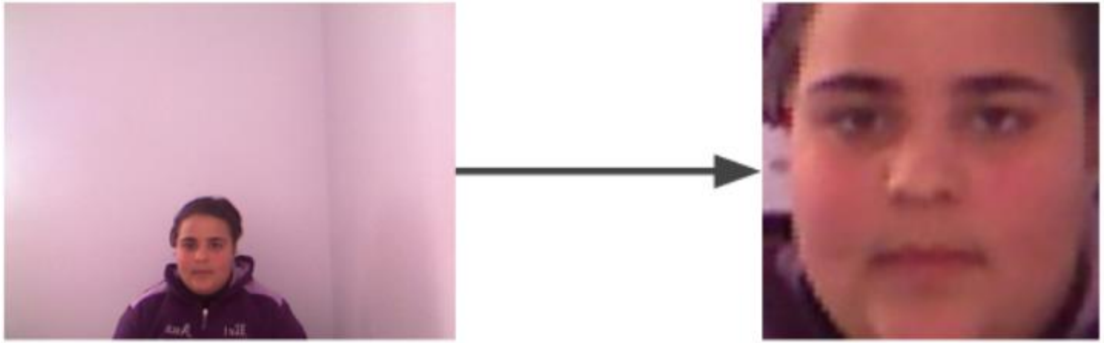
- **MIRACL-VC1:** a lip-reading dataset (Rekik et al., 2014)
  - Captured by Microsoft Kinect sensor, 640x480 pixels
  - 15 speakers (5 men and 10 women)
  - Each speaker read 10 times for a set of 10 words and 10 phrases
  - A total number of 3000 instances (15 x 20 x 10)



ID	Words	ID	Phrases
1	*Begin*	1	*Stop navigation.*
2	*Choose*	2	*Excuse me.*
3	*Connection*	3	*I am sorry.*
4	*Navigation*	4	*Thank you.*
5	*Next*	5	*Good bye.*
6	*Previous*	6	*I love this game.*
7	*Start*	7	*Nice to meet you.*
8	*Stop*	8	*You are welcome.*
9	*Hello*	9	*How are you?*
10	*Web*	10	*Have a good time.*

Data Preprocessing #1

- **Following Garg et al. (2016), Gutierrez & Robert (2017)**
  - Cropped out all but face
  - OpenCV face detection module (Bradski, G., 2000)

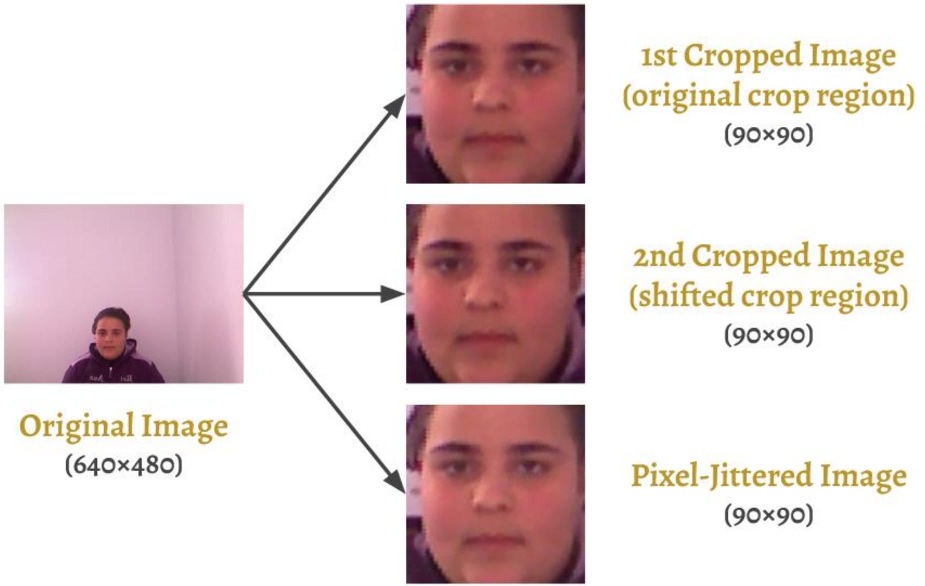


Original Image (640×480)

Cropped Image (90×90)

Data Augmentation

- Following Garg et al. (2016)
  - Tripled dataset size (3,000 → 9,000)
  - 2 modifications: shifting crop region & pixel jittering

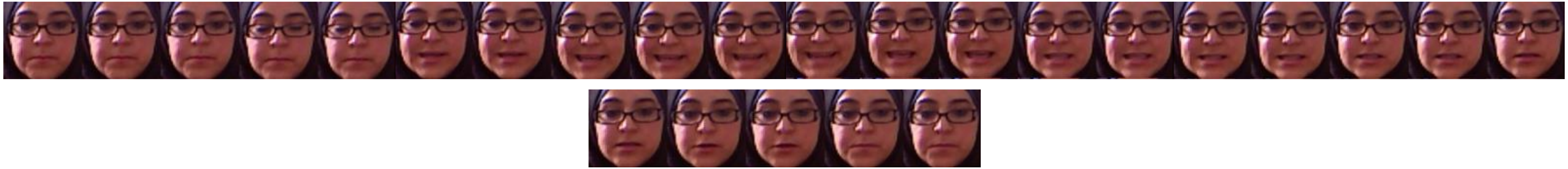


Data Preprocessing #2

- **Each instance:**
  - **Currently:** a sequence of several 90 x 90 pixel images, 1 image/point in time
  - **Desired:** 1 input image/instance
- **Following Garg et al. (2016)**
  - **Step 1:** stretch each sequence

$$stretch\_seq[i] = orig\_seq[floor(\frac{i * orig\_len}{25})]$$

Original Image Sequence for Single Instance (12 90x90 images)



Stretched Image Sequence for Single Instance (25 90x90 images)

Data Preprocessing #2

- **Each instance:**
  - **Currently:** a sequence of several 90 x 90 pixel images, 1 image/point in time
  - **Desired:** 1 input image/instance
- **Following Garg et al. (2016)**
  - **Step 1:** stretch each sequence $stretch\_seq[i] = orig\_seq[floor(\frac{i * orig\_len}{25})]$
  - **Step 2:** concatenate images in stretched sequence

Concatenated Image of Stretched Image Sequence for Same Instance
(one 450x450 image)

Experiment Setup

- Model
  - ViLT (Kim et al., 2021)
    - Pre-tuning: on MSCOCO dataset (200k images)
    - Loss function: cross entropy loss
    - Batch size: 32
    - Epochs: 10 (2250 steps)
- Evaluation
  - For each word/phrase per speaker, 8 for fine-tuning & 2 for testing
    - 6200 instances for fine-tuning & 1800 instances for testing
  - Baselines: random baseline, encoder-decoder approach (CNN + LSTM) (Garg et al., 2016)

Results & Analysis

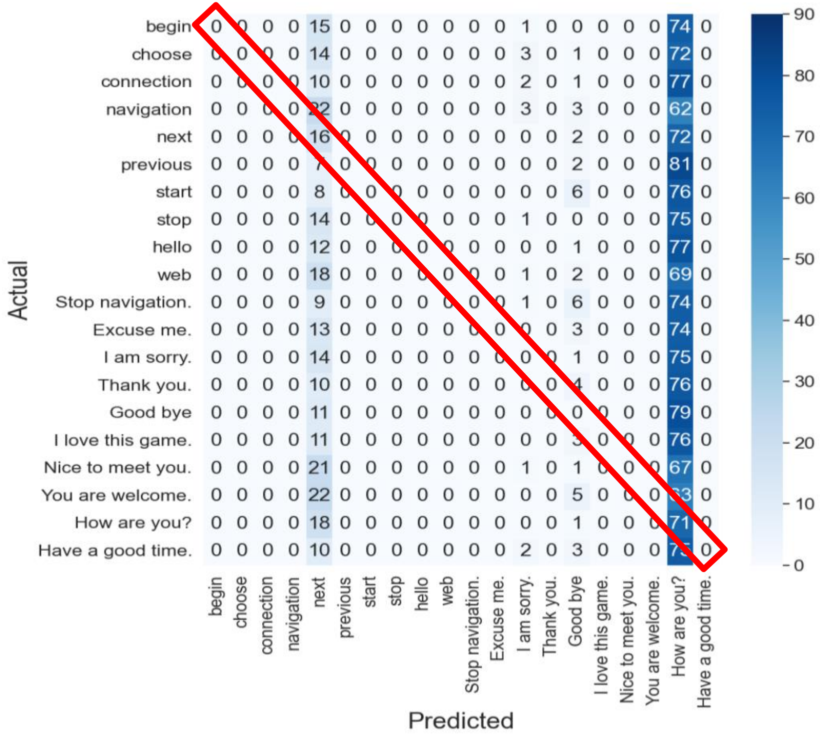
Results Comparison

Table 1: Comparison of testing accuracy among random baseline, CNN and LSTM baseline, ViLT (zero-shot), and ViLT (fine-tuned)

	Only Words	Only Phrases	Both
Random Baseline	10.00%	10.00%	5.00%
CNN + LSTM	56.00%	33.00%	44.50%
ViLT (zero-shot)	7.89%	1.78%	4.83%
ViLT (fine-tuned)	80.44%	98.11%	89.28%

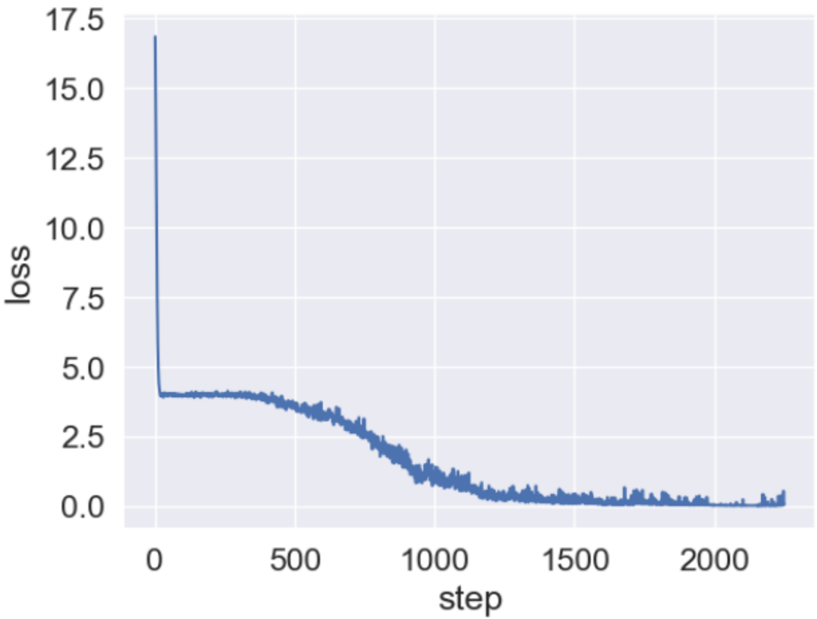
Zero-Shot Results

Heatmap of ViLT zero-shot results

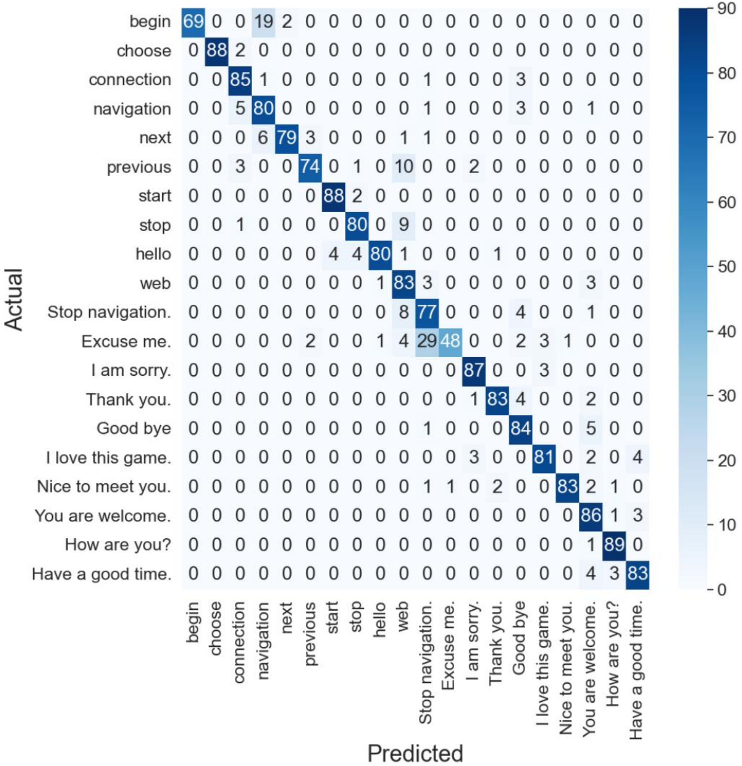


Fine-Tuning Results

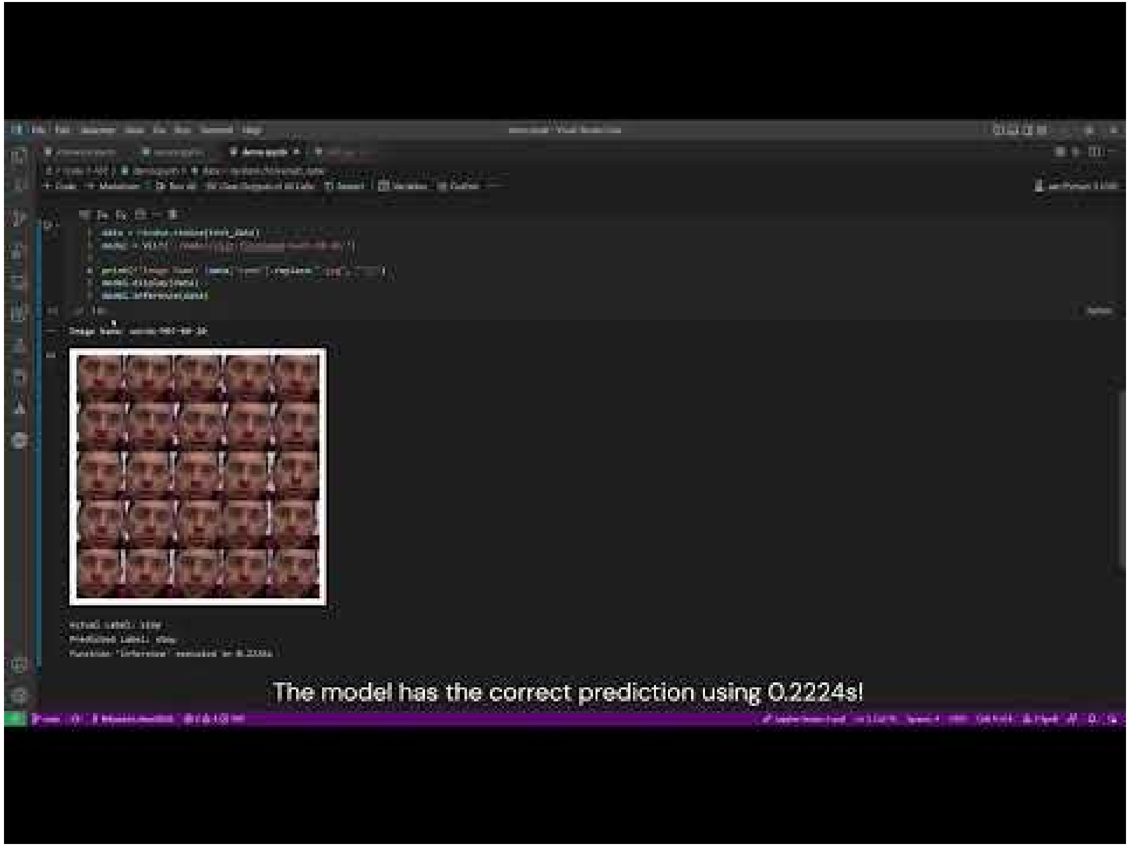
Loss curve while fine-tuning ViLT



Heatmap of ViLT fine-tuned results



Demo Video



Conclusions & Future Works

Conclusions

- Fine-tuned ViLT model can produce promising performance in lip-reading task
  - ~90% overall accuracy and outperformed other baselines
  - ~150 ms inference time

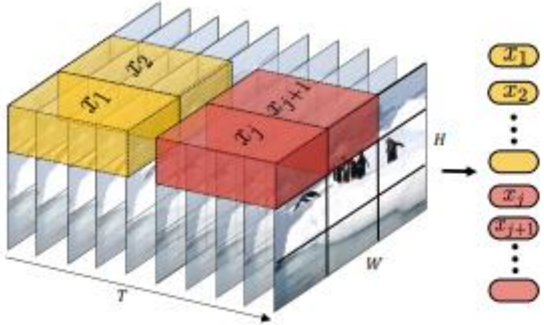
- Multimodal models should be capable for the lip-reading task

- Data preprocessing procedure should be simplified

- Problems with the dataset
  - “Stop”, “Navigation”, and “Stop navigation”
  - Unbalanced gender and skin color distribution

Future Works

- Convert video directly to 3D volume embedding
- Need to check whether ViLT is overfitted
- Use a better dataset
    - More instances
    - More words and phrases
- Investigate other lightweight models
- Deploy the model onto portable devices



(Arnab et al., 2021)

Thank You! Questions?

References

- Abiel Gutierrez and Z Robert. 2017. Lip reading word classification. *Comput Vision-ACCV* (2017).
- Ahmed Rekik, Achraf Ben-Hamadou, and Walid Mahdi. 2014. A New Visual Speech Recognition Approach for RGB-D Cameras. In *Image Analysis and Recognition*, Aurélio Campilho and Mohamed Kamel (eds.). Springer International Publishing, Cham, 21–28. DOI:https://doi.org/10.1007/978-3-319-11755-3_3
- Amit Garg, Jonathan Noyola, and Sameep Bagadia. 2016. Lip reading using CNN and LSTM. *Technical report, Stanford University, CS231 n project report* (2016).
- Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. ViViT: A Video Vision Transformer. DOI:https://doi.org/10.48550/arXiv.2103.15691
- Bradski, G. (2000). The opencv library. Dr. Dobb's Journal of Software Tools.
- Gabriel Loye. 2019. Long Short-Term Memory: From Zero to Hero with PyTorch. FloydHub Blog. Retrieved from https://blog.floydhub.com/long-short-term-memory-from-zero-to-hero-with-pytorch/
- Hao Gu, Yu Wang, Sheng Hong, and Guan Gui. 2019. Blind channel identification aided generalized automatic modulation recognition based on deep learning. *IEEE Access* 7, (2019), 110722–110729.
- Mingfeng Hao, Mutallip Mamut, Nurbiya Yadikar, Alimjan Aysa, and Kurban Ubul. 2020. A Survey of Research on Lipreading Technology. *IEEE Access* 8, (2020), 204518–204544. DOI:https://doi.org/10.1109/ACCESS.2020.3036865
- Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the 38th International Conference on Machine Learning* (Proceedings of Machine Learning Research), PMLR, 5583–5594. Retrieved from https://proceedings.mlr.press/v139/kim21k.html