Exploring End-to-end Attention-based Neural Networks for Native Language Identification

Rutuja Ubale, Yao Qian, Keelan Evanini

Educational Testing Service (ETS), USA

Measuring the Power of Learning.®

Summary

Task Definition

- Native Language Identification (NLI): Automatic identification of the native language (L1) of an individual from their spoken response in a second language (L2).

Motivation

- L1 identification is a challenging research problem that can benefit several spoken language technologies e.g., automatic speech recognition, speaker recognition, interactive voice applications for computer assisted language learning.
- Limited research on the use of spectral features such as MFCC or filter bank features for NLI.
- In end-to-end model, feature representation learning and scoring can be done in a single system.

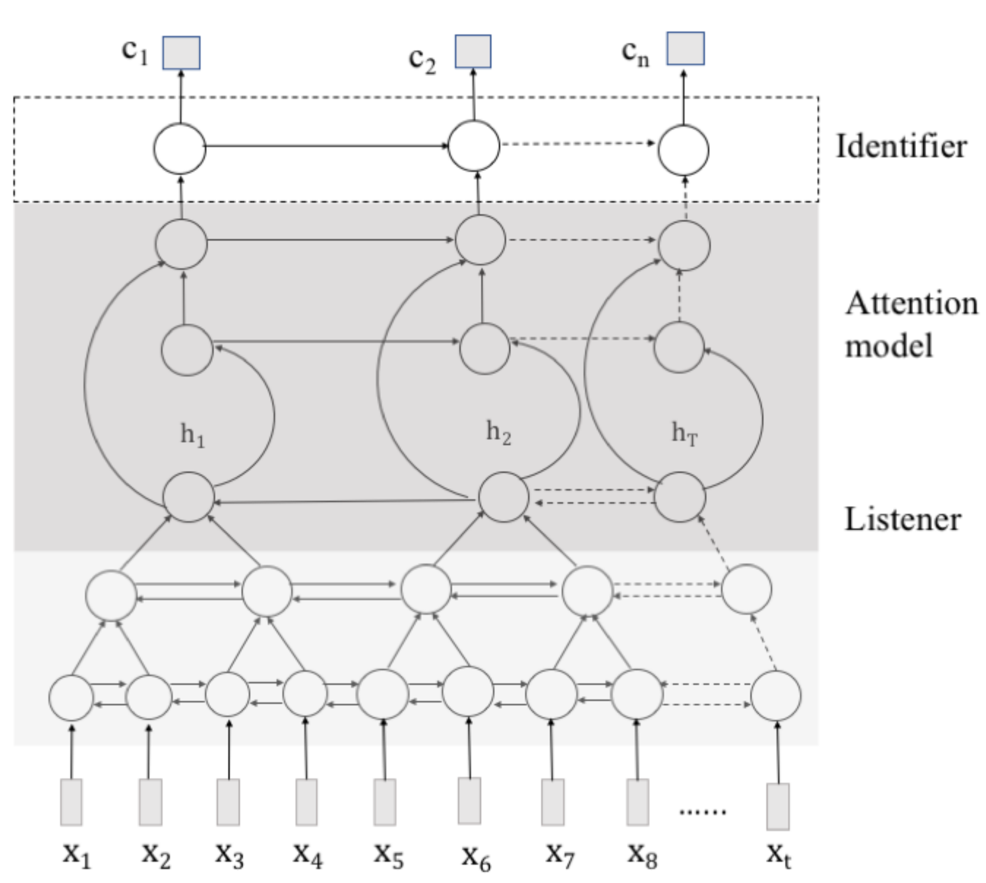
Approach

- Explore different end-to-end architectures for automatic L1 recognition from spectrogram.
  - **Input:** 40 dimensional log-Mel filter bank features
  - **Output:** Posterior probabilities for each L1 class (highest prediction probability is selected as the recognized L1)
- Our end-to-end neural networks consist of three major components:
  - **Encoder network:** Maps input acoustic features to a high-level representation.
  - **Attention model:** Determines which parts in the feature representation are important.
  - **Fully connected classifier network:** Generates a vector of posterior probabilities for each L1.
- Perform score-level fusion using the posterior probabilities generated at the output of the end-to-end models and the log likelihood ratios computed using PLDA scoring model to predict L1.

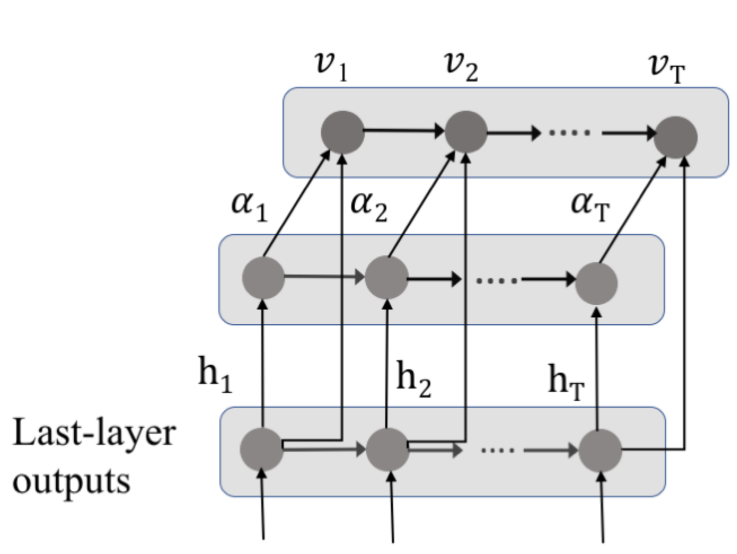
End-to-end Native Language Identification

Listen, Attend and Identify (LAI)

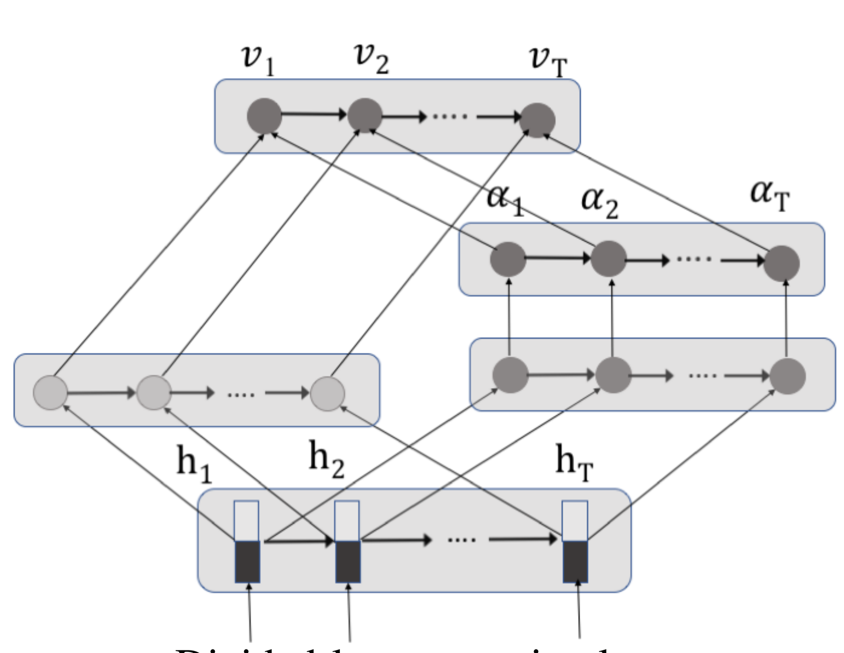
- Encoder Network (Listener) is a three layer pyramidal Bi-GRU (pBGRU) network.
- In every pBGRU, the number of time steps in the feature vector is reduced by one half.
- Attention layer is connected to a fully connected feed-forward layer.



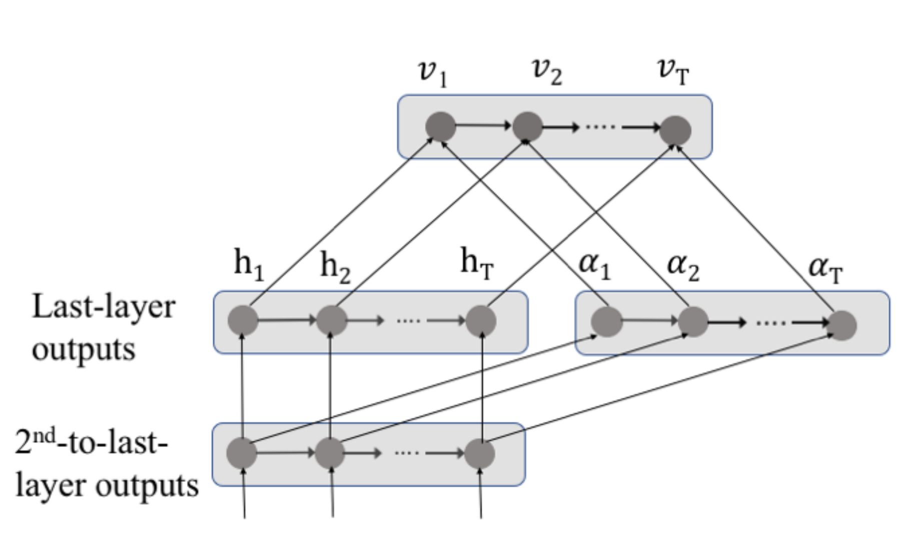
Attention Mechanism



Basic attention layer

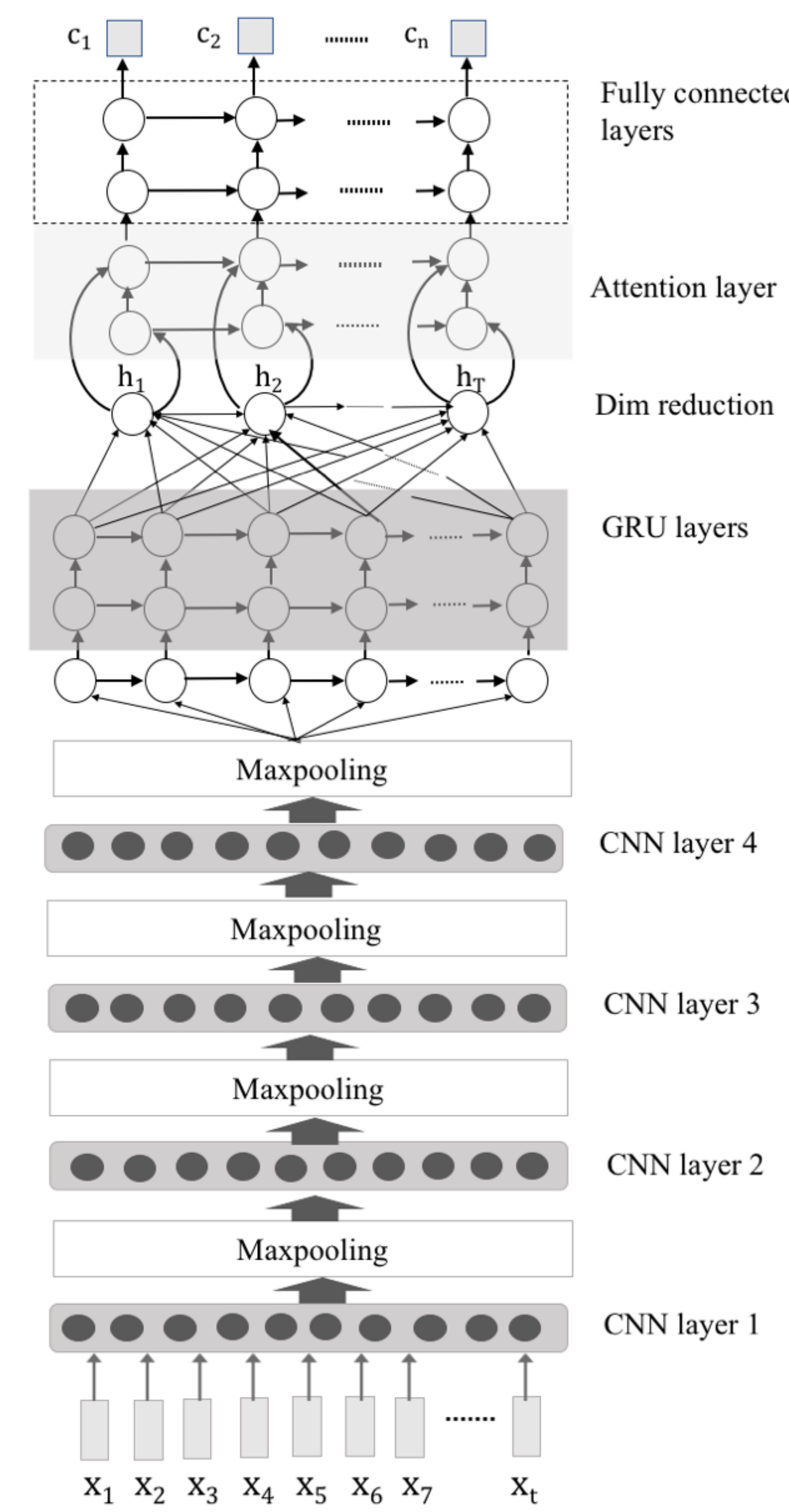


Divided-layer attention layer



Cross-layer attention layer

CGDNN



- Four two-dimensional CNN layers to reduce the frequency variance in the input signal.
- Perform temporal modeling with two uni-directional GRU layers.
- Attention layer is followed by two fully connected DNN layers.

Corpora

- Non-native English speech collected during a high-stakes global assessment of English language proficiency. Each response is approximately 45-60 seconds long.
- 11,000 non-native speakers with 11 different L1 backgrounds: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Spanish, Telugu and Turkish.
- There are approximately 1,000 speech recordings for each L1 in the dataset.

Partition	Train	Validation	Test
Number of audio recordings	7,040	1,760	2,200

Experimental Results

Performance across different NLI systems

Method	Accuracy(%)	UAR(%)
Majority vote baseline	9.00	8.26
RNN only	42.09	42.87
CNN only	60.45	61.13
CGDNN	69.18	69.66
LAI	70.45	70.87
i-vector baseline	79.72	81.59

Performance with different attention layers

Model	Basic	Cross-layer	Divided-layer
LAI	70.45	71.72	68.63
CGDNN	69.18	70.18	69.09

Performance across different fusion systems

Fusion system	Basic	Cross-layer
LAI + i-vector	82.13	82.27
CGDNN + i-vector	82.86	83.14
LAI + CGDNN + i-vector	83.18	83.32

- Our best attention-based neural network can achieve a performance approaching the performance of the i-vector system.
- Fusion of the end-to-end system with the i-vector system leads to significant performance improvements, indicating that the three systems are able to capture complementary information from the data.