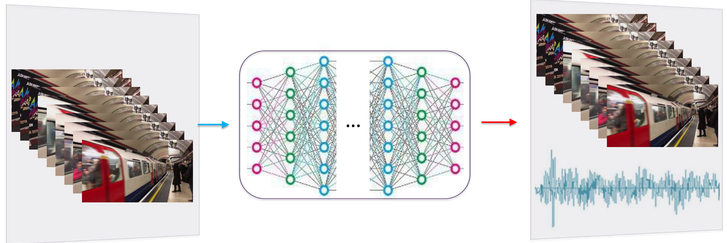
Kehinde Olufowobi, MS in Computer Science
Advisor: Dr. Nic Herndon

Visual and Acoustic Relationships in Large-Scale Video Understanding

What

Imagine a video clip with no sound. Now imagine feeding that clip into an algorithm that reproduces the exact video along with artificially-generated matching sound realistic enough to fool a human.



Silent video — Deep neural net generator — Video with realistic sound

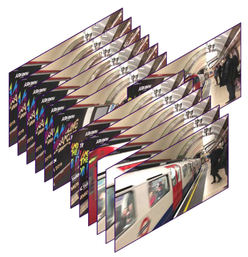
I present an ongoing investigation that involves employing deep neural networks to model physical interactions within a visual scene towards synthesizing realistic sound from silent visuals.

How

1. Start with a diverse, large-scale dataset:

6.1 million videos, 350 thousand hours, 2.6 billion audio/visual features

Each video comprises of 300 subsampled frames spanning 1 second each. Corresponding to each frame is a 1024-dimensional RGB representation and a 128-dimensional audio representation.

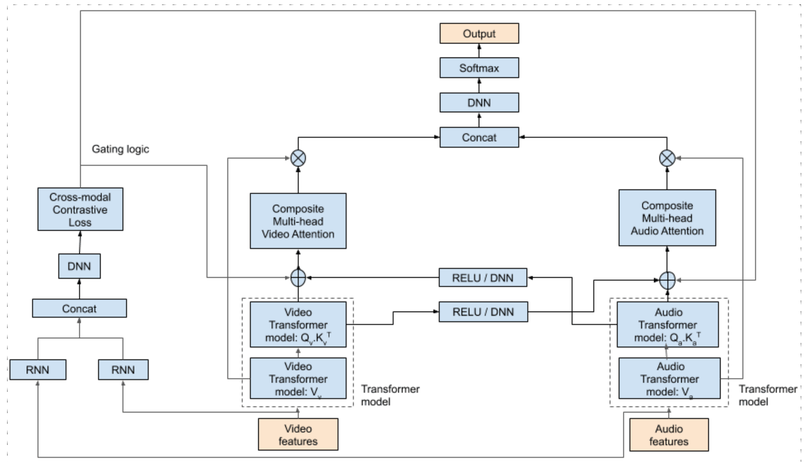


2.
Refine the dataset by selecting the subset of videos with sufficiently high or sufficiently low audio/video correlation using a cross-modal Transformer model with a cross-entropy-based correlation tower [1] (trained on positive and negative examples of audio/video correlation).

3.
Formulate the task as a conditional generation problem for which a conditional generative model is trained to synthesize raw waveform samples from an input video. Estimate the following conditional probability:

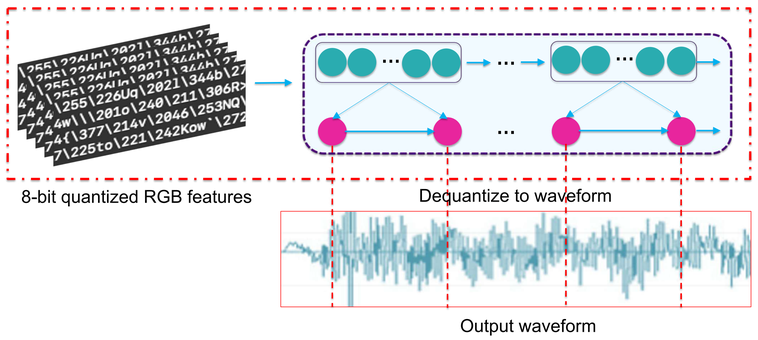
$$p(y_1, y_2, ..., y_n | x_1, x_2, ..., x_m)$$

where x_1 ... x_m are input video frames representations and y_1 ... y_n are output waveform values.



[1] A cross-modal Transformer with correlation tower

4.
Using the refined dataset, employ a sequence-to-sequence model to synthesize sound from video. Use the video (RGB) representation to initialize the hidden state (green nodes in [2] below) of the coarsest-tier recurrent neural net (RNN) of a sound generator.



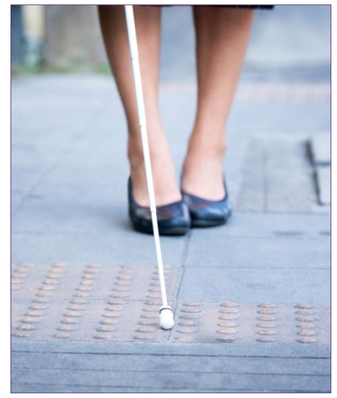
[2] Sequence-to-sequence model architecture

5.
Qualitative Evaluation:
Visualize and compare the generated waveform results to the real audio, then listen to the generated sound and the real audio.
Quantitative Evaluation:
Assess the average cross-entropy loss for model training and testing. Then conduct a retrieval experiment where visual features are used as queries and the audio with the maximum sampling likelihood is retrieved.
Human Evaluation:
Enlist human participants with significantly diverse backgrounds to assess the synchronization between video and audio and to assess the quality of the model result via a real-or-fake discrimination task.

Keywords: representation learning, deep learning, audio-visual scene understanding

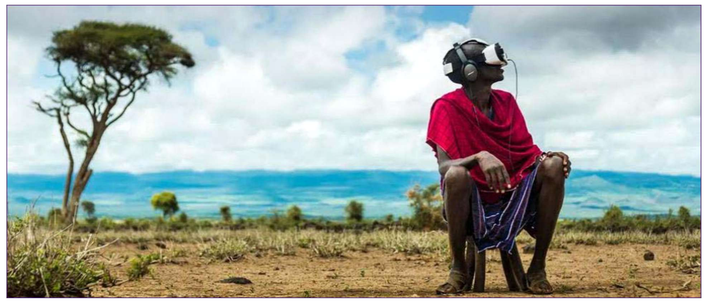
Why

We can enable/automate/enhance:

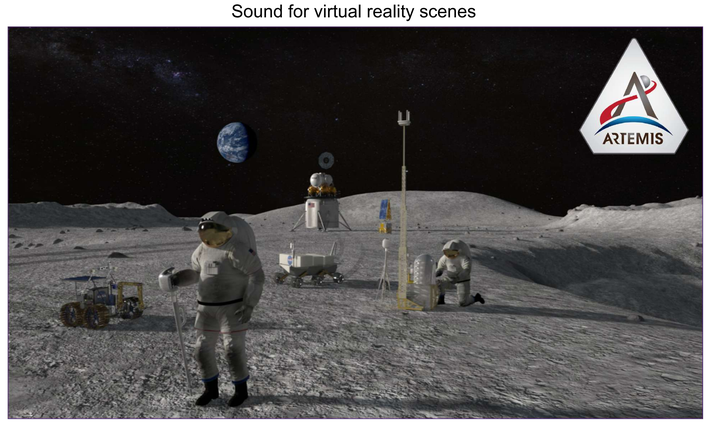


Accessibility for the visually impaired

Foley in filmmaking



Sound for virtual reality scenes



Hyperrealistic acoustic awareness for astronauts in the vacuum of space

References