Jointly discovering Visual Objects and Spoken Words from Raw Sensory Input

- Video Presentation at https://www.youtube.com/watch?v=fNm4fh2ub9c

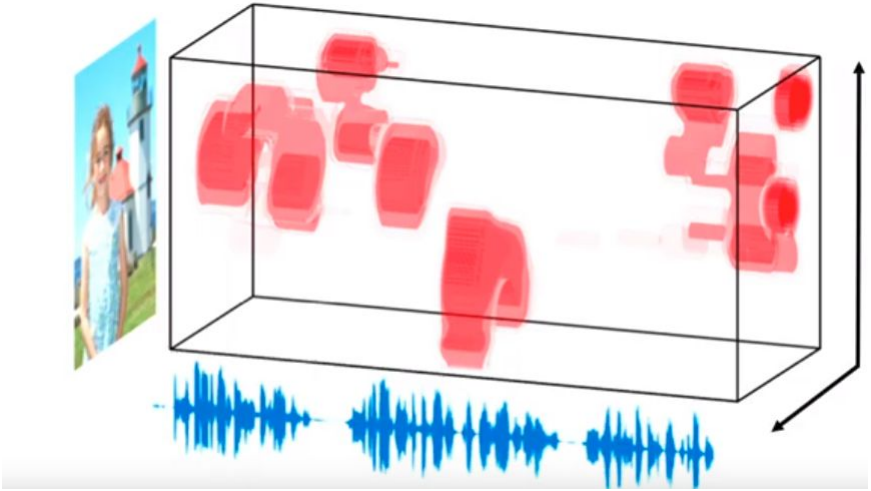
Objectives

- Let's take inspiration from how babies learn things.

- Obtain segmentation of objects in the image from raw pixels.

- Obtain segmentation of words in speech from raw audio.

- Obtain the above jointly such that visual-acoustic associations are preserved

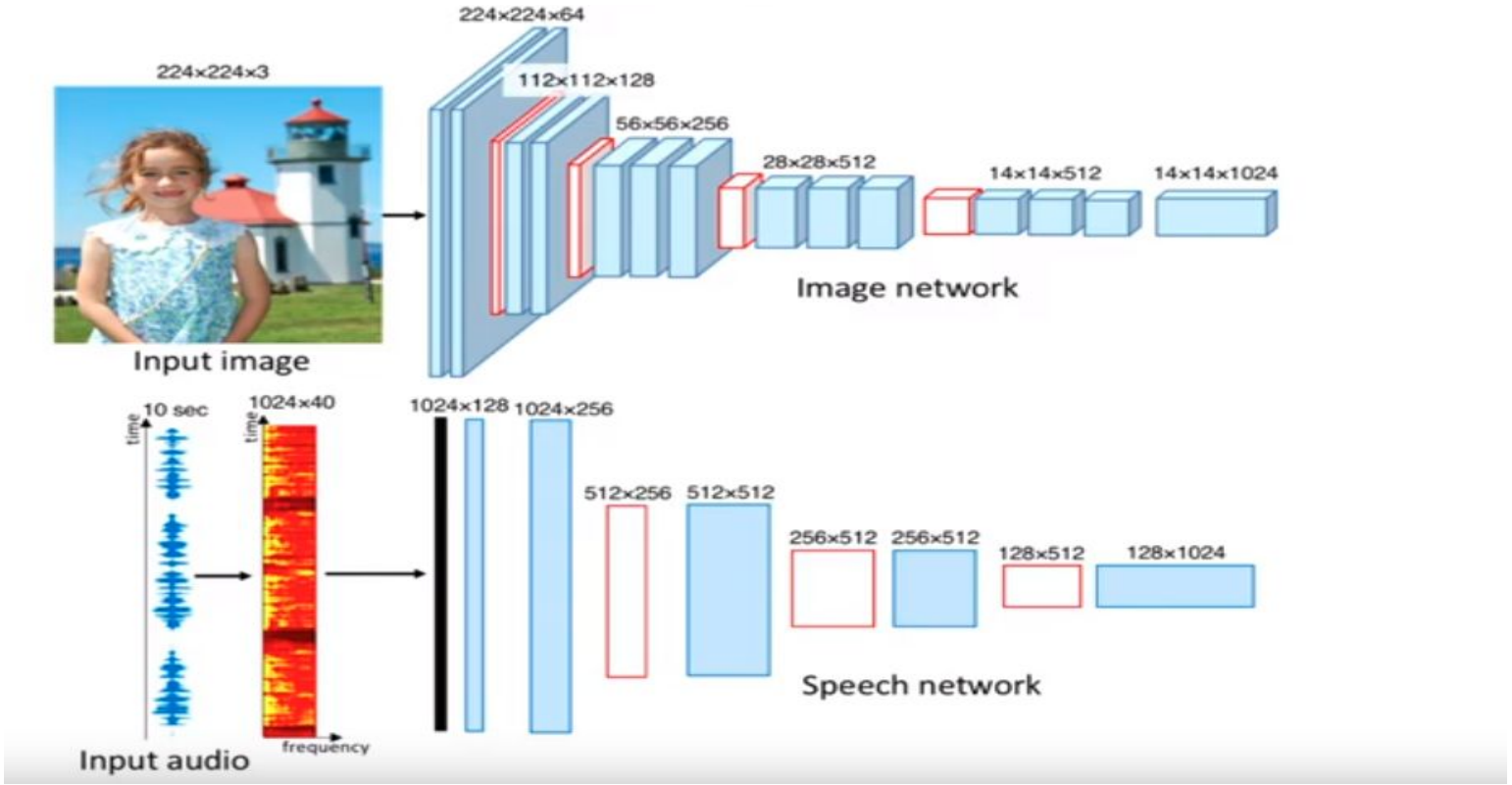


Datasets

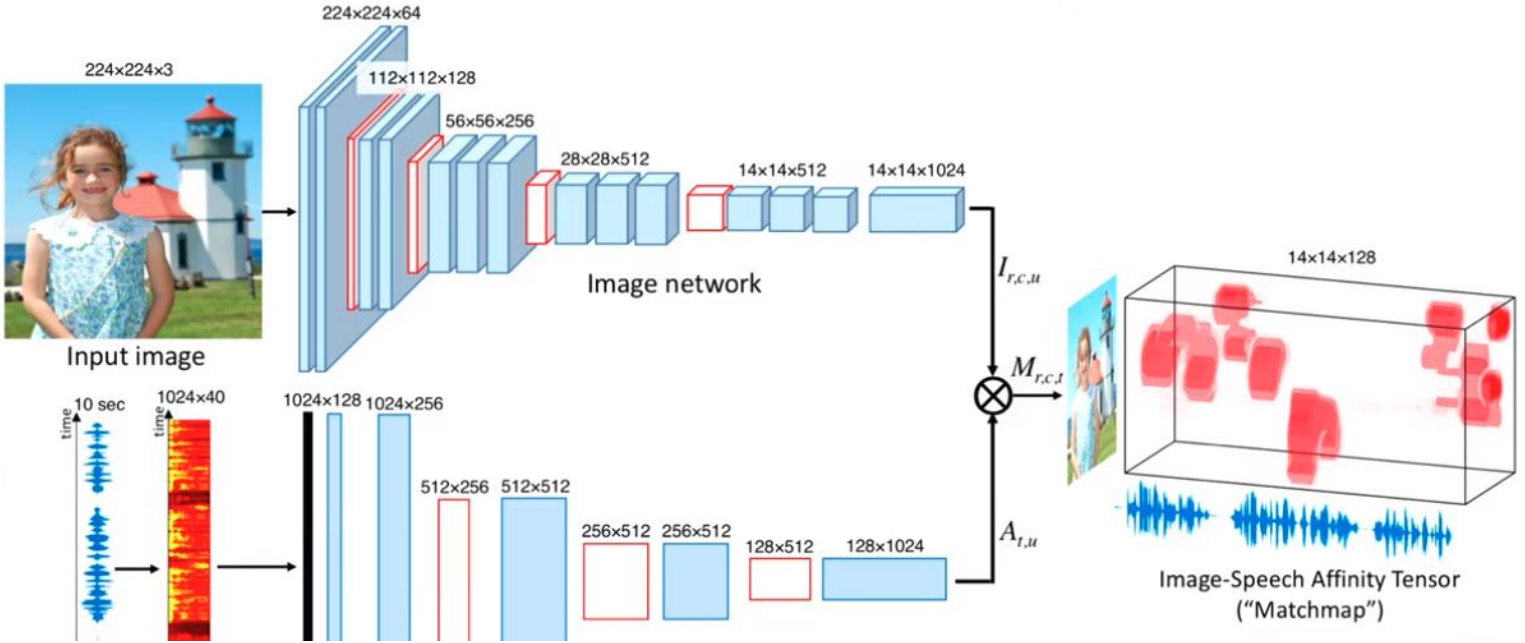
- Places Audio Captions
- Crowd sources
- Describe the image to a blind person.

- 200K audio captions (+ augment by 200K more) -> 402385 pairs for training + 1000 for validation
- Vocab size of 44,342.
- 2683 speakers

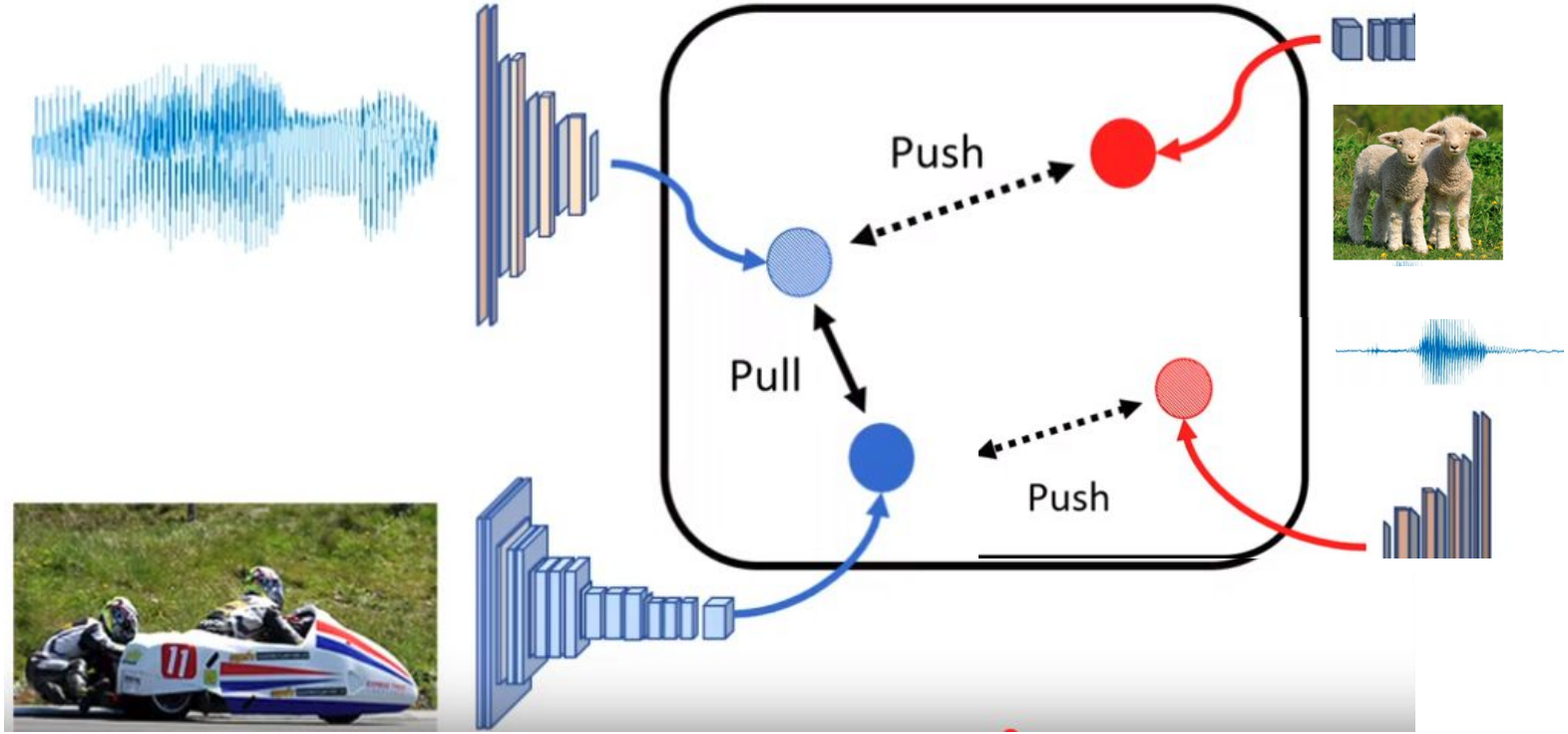
Unimodal representations learnt via Deep Audio Visual Embedding Network (DAVE Nets)



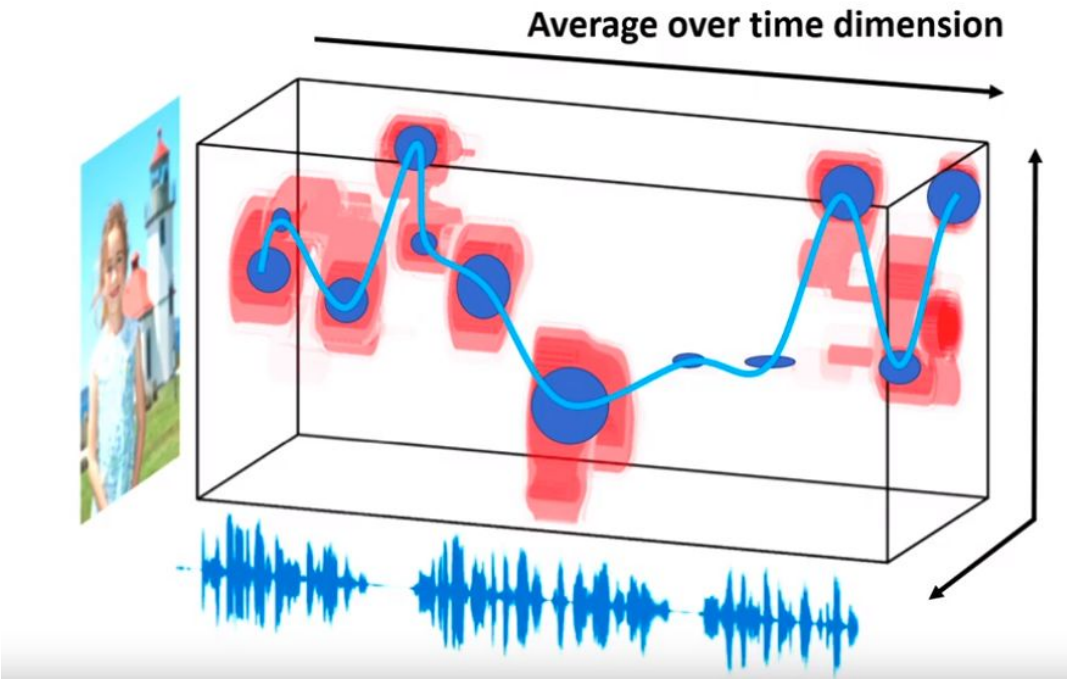
Unimodal representations fused to obtain spatio temporal shared representations



Model is trained to localize using Triplet Loss



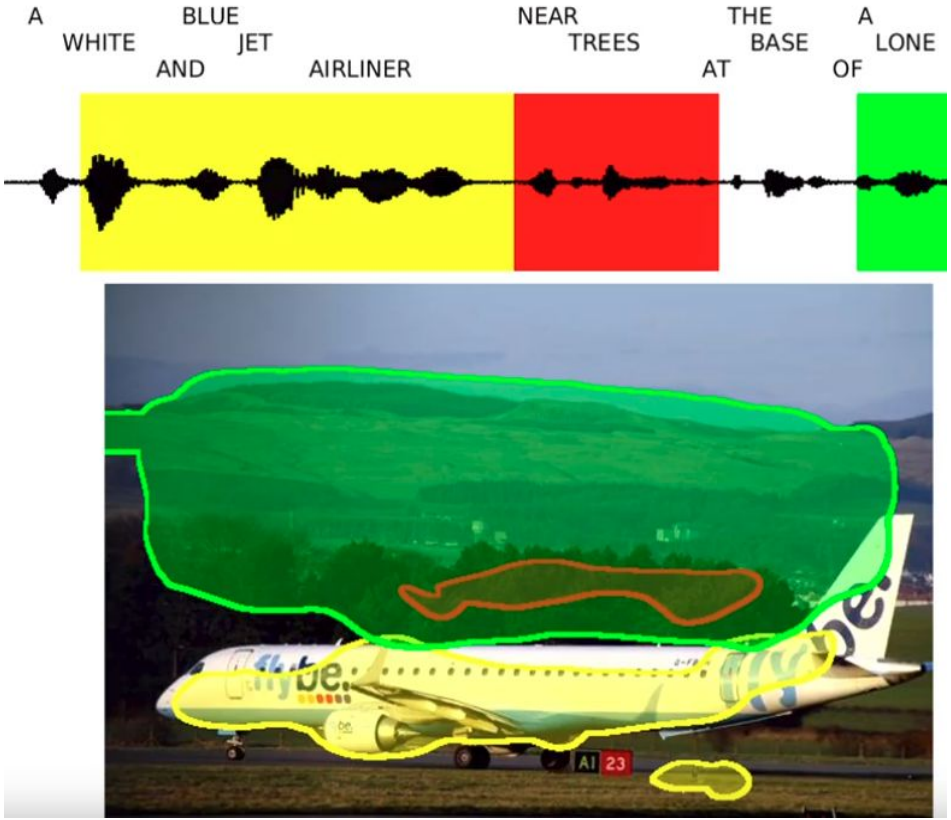
Pooling is performed to obtain a trajectory across time



Hyperparameters:

- SGD 0.001 + 0.9 M
- Batch size 128
- LR decay by 10 every 70 epochs
- Convergence ~150 epochs

Audio visual associative localizations seem to emerge from training



What if we ping a single neuron?