# Multimodal Deep Learning

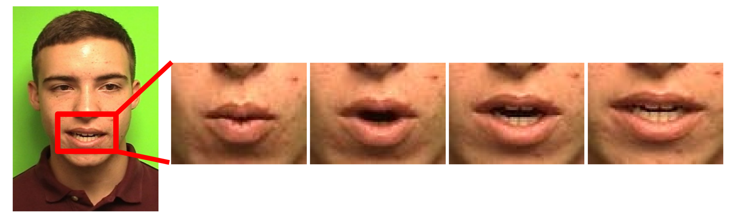
Jiquan Ngiam[1], Aditya Khosla[1], Mingyu Kim[1], Juhan Nam[1], Honglak Lee[2] & Andrew Ng[1]

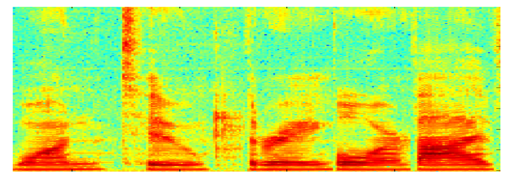
[1]Stanford University, [2]University of Michigan, Ann Arbor

## 1 Overview

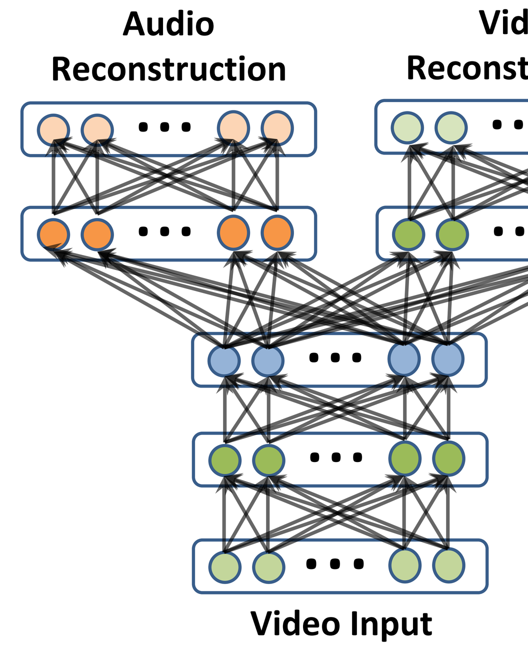
- Most deep learning methods have been to applied to only *single* modalities (single input source).
- A straightforward approach to multimodal data (multiple input sources) is ineffective.
- We propose novel deep architectures for learning over multimodal data that effectively learn to relate audio and video data.

- Data: Video recordings of subjects saying digits and letters
- Task: Audio-visual speech classification
- Key Challenges:
  - **Cross Modality Learning**: If our task is visual-only recognition (lipreading), can we learn better video features by using audio to adapt the features?
  - **Multimodal Feature Learning**: Designing multimodal features is difficult; can we learn multimodal features that integrate audio and visual information?

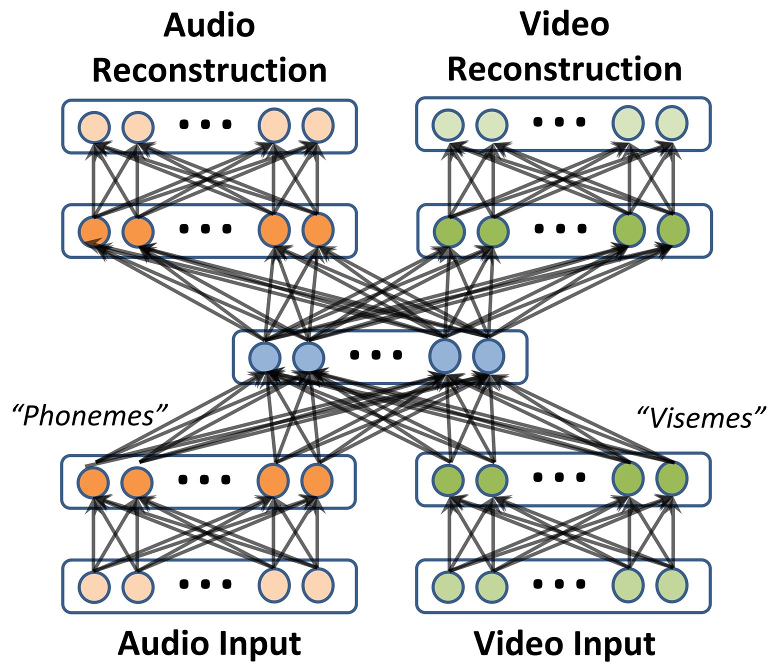




## 2 Multimodal Deep Network Architectures




**Video-only Deep Autoencoder (Cross-modality Learning)**




**Bimodal Deep Autoencoder**

- Learns video representations that try to reconstruct audio (from audio-video pairs of examples)
- Since audio works well for speech recognition, this discovers good video representations for visual speech recognition (lip-reading)

- Trained with "hidden data" – reconstruct both outputs given only one (e.g. video-only) input: 1/3 of data requires the model to reconstruct both audio and video, given only video input; 1/3 with only audio input; 1/3 with both inputs

## 3 Visual Speech Recognition (Lip-reading)

| Feature Learning | Supervised Learning | Testing |
|---|---|---|
| Audio + Video | Video | Video |

| Feature Representation | Accuracy |
|---|---|
| Baseline "Raw" Video Input | 46.2% |
| Video-Only Learning (Single Modality Learning) | 54.2% ± 3.3% |
| **Our Features (Cross Modality Learning)** | **64.4%** ± 2.4% |
| Multiscale Spatial Analysis [1] | 44.6% |
| Local Binary Pattern [2] | 58.9% |

AVLetters Performance (26-way Classification)

By learning better video features using audio as a cue (video-only deep autoencoder), we are able to achieve performance superior to best published results on AVLetters.

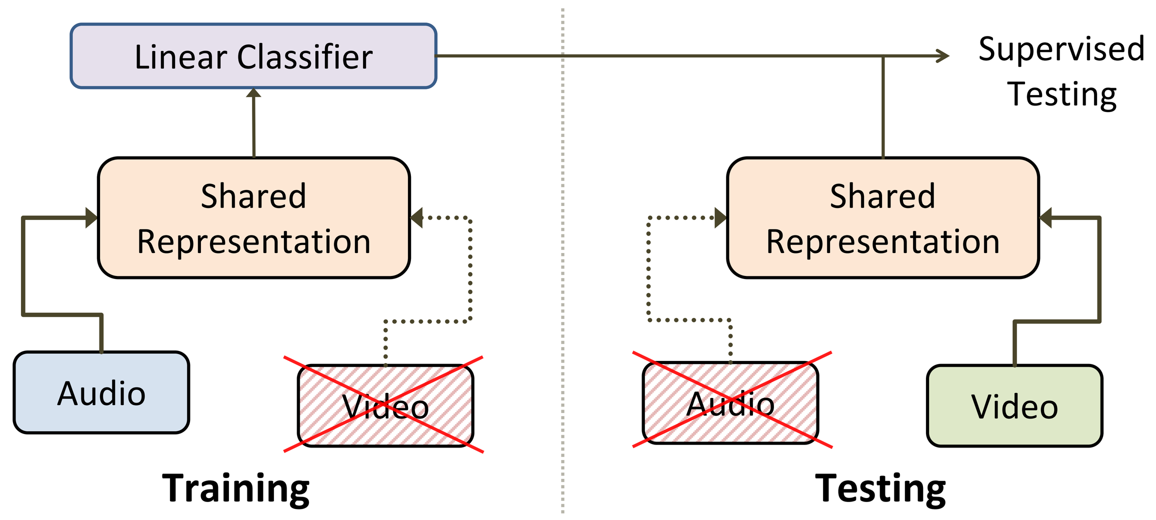
| Feature Representation | Accuracy |
|---|---|
| Baseline "Raw" Video Input | 58.5% |
| Video-Only Learning (Single Modality Learning) | 65.4% ± 0.6% |
| **Our Features (Cross Modality Learning)** | **68.7% ± 1.8%** |
| Discrete Cosine Transform [3] | 64% |
| Active Appearance Model [4] | 75.7% |
| Active Appearance Model [5] | 68.7% |
| Fused Holistic + Patch [6] | 77.1% |
| Visemic AAM [7] | 83% |

CUAVE Performance (10-way Classification)

We also see an improvement when using audio as a cue on the CUAVE dataset.

## 4 Shared Representation Learning

| Feature Learning | Supervised Learning | Testing |
|---|---|---|
| Audio + Video | Audio | Video |
| Audio + Video | Video | Audio |




| Train / Test | Method | Accuracy |
|---|---|---|
| Train Audio, Test Video | Raw Data + CCA | 41.9% |
| | Learned Features + CCA | **57.3%** |
| Train Video, Test Audio | Raw Data + CCA | 42.9% |
| | Learned Features + CCA | **91.7%** |

- Use canonical correlation analysis (CCA) to learn a linear map that forms a shared representation between the audio and video modalities
- Learned features + CCA does surprisingly well to find a shared representation between audio and video

## 5 Multimodal Fusion

| Feature Learning | Supervised Learning | Testing |
|---|---|---|
| Audio + Video | Audio + Video | Audio + Video |

| Feature Representation | Accuracy (Clean Audio) | Accuracy (Noisy Audio) |
|---|---|---|
| Learned Audio Features (RBM) | **95.8 %** | 79.6 % |
| Learned Video Features | 68.7 % | 68.7 % |
| Bimodal Deep Autoencoder | 90.0 % | 77.6 % ± 1.4% |
| Learned Video Features + Audio Features | 87.0 % | 76.6 % ± 0.8% |
| **Bimodal Features + Audio Features** | 94.4 % | **82.2 % ± 1.2%** |

- Fusing audio features and bimodal features can improve performance over audio-only features, especially when the audio is degraded with noise.

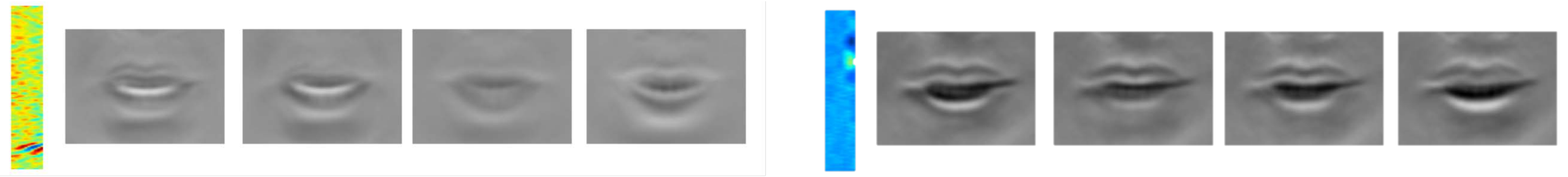
## 6 Simulating the McGurk Effect

| Feature Learning | Supervised Learning | Testing |
|---|---|---|
| Audio + Video | Audio + Video | Audio + Video |

- The McGurk effect is an audio-visual perception phenomenon where a visual */ga/* with an audio */ba/* is perceived as */da/* by most subjects.
- We collected data from volunteers saying */ga/*, */ba/* and */da/* for a three-way classification task.
- Our model reflects the same perception phenomenon.

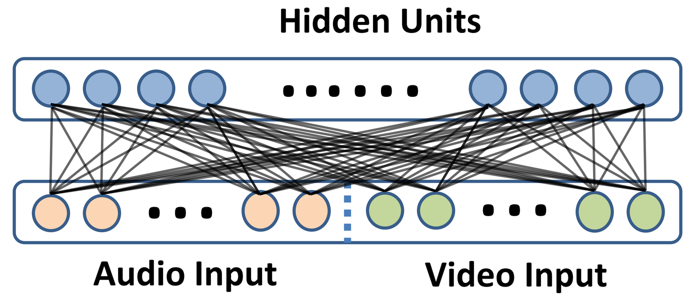
| Audio + Visual Setting | Model Predictions | | |
|---|---|---|---|
| | */ga/* | */ba/* | */da/* |
| Visual */ga/* + Audio */ga/* | 82.6% | 2.2% | 15.2% |
| Visual */ba/* + Audio */ba/* | 4.4% | 89.1% | 6.5% |
| Visual */ga/* + Audio */ba/* | 28.3% | 13.0% | **58.7%** |

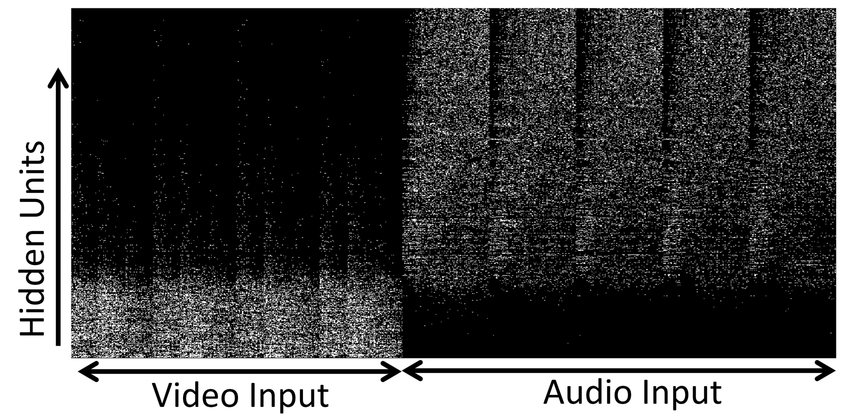
## 7 Visualizing Learned Features



- We learn video features (e.g. showing of teeth, capturing mouth motion) that can help determine the place of articulation.
- The deep hidden units also learn to relate video features to audio features.

## 8 Control Experiments




Weight matrix learned by a shallow RBM

- A shallow RBM tends to learn hidden units that are strongly connected to either modality and few that connect across

## 9 References

[1] I. Matthews, T.F. Cootes, J.A. Bangham, and S. Cox. Extraction of visual features for lipreading. PAMI, 24, 2002.
[2] G. Zhao and M. Barnard. Lipreading with local spatiotemporal descriptors. IEEE Transactions on Multimedia,11(7):1254–1265, 2009.
[3] M. Gurban and J.P. Thiran. Information theoretic feature extraction for audio-visual speech recognition. IEEE Transactions on Signal Processing, 57(12):4765–4776, 2009.
[4] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos. Multimodal fusion and learning with uncertain features applied to audiovisual speech recognition. In MMSP, 2007.
[5] V. Pitsikalis, A. Katsamanis, G. Papandreou, and P. Maragos. Adaptive multimodal fusion by uncertainty compensation. In ICSLP, pages 2458–2461, 2006.
[6] P. Lucey and S. Sridharan. Patch-based representation of visual speech. In HCSNet Workshop on the Use of Vision in HCI, 2006
[7] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. IEEE TASLP, 2009.

http://ai.stanford.edu/~jngiam/