CAN DNNs LEARN TO LIPREAD FULL SENTENCES?

George Sterpu, Christian Saam, Naomi Harte

Trinity College Dublin, Ireland
Lipreading in the news

Intelligent Machines

AI Has Beaten Humans at Lip-reading

A pair of new studies show that a machine can understand what you're saying without hearing a sound.

by Jamie Condliffe  November 21, 2016

Emergent Tech  Artificial Intelligence

Is that you, HAL? AI can now see secrets through lipreading – kinda

LipNet's got potential but also a looooong way to go

By Katyanna Quach 8 Nov 2016 at 00:59

Can deep learning help solve lip reading?

New research paper shows AI easily beating humans, but there's still lots of work to be done

By James Vincent | @jvinct | Nov 7, 2016, 12:50pm EST
Human level performance

Word Accuracy - Human vs AI

- **GRID**: 28 hours laboratory LipNet
- **LRS**: 246 hours “in the wild” WLAS
- **LSVSR**: 3,886 hours “in the wild” LipNet 2
What is lip-reading?
What is lip-reading?
Bad lip-reading?
Lip-reading strategies

**Prior work**

“end to end”
- collect lots of data
- train a large DNN
- decode *text*

“traditional”
- simpler task
- DCT (etc.) + HMM
- decode *visemes*

**Our work**

“best of both worlds”
- large vocabulary
- full sentences
- train a small DNN
- decode *visemes*
Visemes = unambiguous units

Phonemes | Word Examples
---|---
ch, sh, jh, zh | cheap, sheep

Phonemes | Word Examples
---|---
m, em, b, p | mat, bat, pat

Phonemes | Word Examples
---|---
f, v | fail, veil

TCD-TIMIT Dataset

<table>
<thead>
<tr>
<th>speakers</th>
<th>62</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentences</td>
<td>6619</td>
</tr>
<tr>
<td>size</td>
<td>9 hours</td>
</tr>
<tr>
<td>vocabulary</td>
<td>6224 words</td>
</tr>
<tr>
<td>resolution</td>
<td>1920x1080</td>
</tr>
</tbody>
</table>

Publicly available: sigmedia.tcd.ie/TCDTIMIT

A few examples:
- he took his mask from his forehead and threw it unexpectedly across the deck
- civilization is what man has made of himself
- the mayan neoclassic scholar disappeared while surveying ancient ruins
CNN + Seq2seq architecture

CTC Loss

Connectionist Temporal Classification (CTC)

Memory

LSTM
LSTM
LSTM
LSTM

Convolutional Neural Network (CNN)

CTC Loss

Sequence Loss

Softmax Cross-Entropy

Attention (soft selection)

Summary

Video files

O₁ O₂ O₃ O₃

Y₁ Y₂ Y₃ END

LSTM LSTM LSTM LSTM

GO Ŷ₁ Ŷ₂ Ŷₙ
Viseme recognition accuracy [%]
Viseme recognition accuracy [%]
Viseme recognition accuracy [%]

- DCT + HMM
- AAM + HMM
- EigenLips + DNN-HMM
- zeros + LSTM
- DCT + LSTM
- DCT + BiLSTM
- DCT + LSTM w/ attention
- DCT + LSTM + monotonic attn
- 2D CNN + BiLSTM
- 2D CNN + LSTM (CTC-CE)
- 2D ResNet + LSTM
- 2D CNN + LSTM
- 3D CNN + LSTM
- 64x64 2D CNN + LSTM (CTC-CE)
- 3D CNN + LSTM
Viseme recognition accuracy [%]
Viseme recognition accuracy [%]

The chart shows the viseme recognition accuracy for various methods. The methods are listed on the x-axis and the accuracy is represented on the y-axis. The methods include:

- DCT + HMM
- AAM + HMM
- Eigens + DNN-HMM
- zeros + LSTM
- DCT + LSTM
- DCT + BiLSTM
- DCT + LSTM w/ attention
- DCT + LSTM + monotonic attn
- DCT + LSTM (CTC-CE)
- 2D CNN + BiLSTM
- 2D CNN + LSTM (CTC-CE)
- 2D ResNet + LSTM
- 2D CNN + LSTM
- 3D CNN + LSTM
- 64x64 2D CNN + LSTM
- 2D CNN + LSTM

The highest accuracy is achieved by the 2D CNN + LSTM (CTC-CE) method, with a recognition accuracy of approximately 75%.
Conclusion

Our work
“Visual Sentence to Visemes DNN”

● Pros:
  ○ easier to train

● Cons:
  ○ manually defined visemes
  ○ not for Video-only applications

● Potential:
  ○ Audio-Visual fusion
  ○ automatically learn visual speech units
● Our code is publicly available on Github
  🐦 github.com / georgesterpu / Sigmedia-AVSR

■ TensorFlow-based seq2seq network for
  **Speech Recognition**
  ○ Audio only
  ○ Visual only
  ○ Audio-Visual fusion

@John_Tukey