LipNet
End-to-End Sentence-level Lipreading

Yannis Assael, Brendan Shillingford, Shimon Whiteson, Nando de Freitas
Outline

1. Introduction
2. Background
3. LipNet
4. Analysis
How easy do you think lipreading is?

- McGurk effect (McGurk & MacDonald, 1976)
- Phonemes and Visemes (Fisher, 1968)
- Human lipreading performance is poor

We can improve it…
1. Introduction

Sentence: Place blue in m 1 soon
LipNet:

https://goo.gl/hyFBVQ
Why is lipreading important?

Among others:

- Improved hearing aids
- Speech recognition in noisy environments (e.g. cars)
- Silent dictation in public spaces
- Security
- Biometric identification
- Silent-movie processing
1. Introduction

https://goo.gl/RTXh9Q
Automated lipreading

- Most existing work does not employ deep learning
- Heavy preprocessing
- Open problems:
  - generalisation across speakers
  - extraction of motion features
End-to-end supervised learning using NNs

1. Hierarchical, expressive, differentiable function

1. Adjust parameters to maximise probability of data with gradient descent
Convolutional Neural Networks

- **Model:** Deep stacks of local operations.
- **Good for:** relationships over **space (2D):**

Also good for **time (1D)**

Or in our case, **space & time (3D):** every layer can model either or both. Let's the optimisation decide what's best.
Recurrent Neural Networks

- **Model**: carry information over time using a state
- **Good for**: sequences

- Often used to predict classes at each timestep
- But what if inputs/outputs are unequal length, or aren't aligned?
Recurrent Neural Networks

- If inputs/outputs aren't aligned, CTC (Graves 2006) efficiently marginalises over all alignments

- To do this, let the RNN output **blanks** or **duplicates**:

\[
p(am) = p(aam) + p(amm) + p(_am) + p(a_m) + p(am_)
\]

- Sum over every way to output the same sequence:
LipNet

- Monosyllabic vs Compound words (Easton & Basala, 1982)
- Spatiotemporal features
- End-to-end, sentence-level
- GRID corpus 33000 sentences

**TABLE I. Sentence structure for the Grid corpus. Keywords are identified with asterisks.**

<table>
<thead>
<tr>
<th>command</th>
<th>color*</th>
<th>preposition</th>
<th>letter*</th>
<th>digit*</th>
<th>adverb</th>
</tr>
</thead>
<tbody>
<tr>
<td>bin</td>
<td>blue</td>
<td>at</td>
<td>A–Z</td>
<td>1–9, zero</td>
<td>again</td>
</tr>
<tr>
<td>lay</td>
<td>green</td>
<td>by</td>
<td>excluding W</td>
<td>now</td>
<td></td>
</tr>
<tr>
<td>place</td>
<td>red</td>
<td>in</td>
<td></td>
<td>please</td>
<td></td>
</tr>
<tr>
<td>set</td>
<td>white</td>
<td>with</td>
<td></td>
<td>soon</td>
<td></td>
</tr>
</tbody>
</table>
GRID corpus

Sentence: Place blue in m 1 soon
LipNet:

Sentence: Set green with u zero again
LipNet: Set green with u zero again

Sentence: Lay red in m 3 soon
LipNet: Lay red in m 3 soon
Preprocessing

- Facial Landmarks
- Crop the mouth
- Affine transform the frames
- Smoothen using Kalman filter
- Temporal augmentation
Model Architecture

- t frames
- STCNN + Spatial Pooling (x3)
- Bi-GRU (x2)
- Linear
- CTC loss
Baselines

- Hearing-Impaired People
  3 students from the Oxford Students’ Disability Community

- Baseline-LSTM
  Replicate previous state-of-the-art architecture by (Wand et al., 2016)

- Baseline-2D
  Spatial-only convolutions

- Baseline-NoLM
  Language model disabled
## Lipreading Performance

<table>
<thead>
<tr>
<th></th>
<th>Unseen Speakers</th>
<th>Overlapped Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CER</td>
<td>WER</td>
</tr>
<tr>
<td>Hearing Impaired</td>
<td>47.7%</td>
<td></td>
</tr>
<tr>
<td>Baseline-LSTM</td>
<td>38.4%</td>
<td>52.8%</td>
</tr>
<tr>
<td>Baseline-2D</td>
<td>16.2%</td>
<td>26.7%</td>
</tr>
<tr>
<td>Baseline-NoLM</td>
<td>6.7%</td>
<td>13.6%</td>
</tr>
<tr>
<td>LipNet</td>
<td>6.4%</td>
<td>11.4%</td>
</tr>
</tbody>
</table>
Learned Representations

4. Analysis
Viseme Confusions

(a) Lip-rounding vowels  (b) Alveolar  (c) Bilabial  (d) Viseme Categories
Thank you!
Thank you NVIDIA!