Deep Lip Reading: a comparison of models and an online application

January 20, 2021

Outline

1 Context & Motivation

2 LipNet

3 Deep Lip Reading

Lip reading: what is it and what role does it play?

- The **ability** to **recognize what** is **being** said based on **visual information**

Lip reading: what is it and what role does it play?

- The **ability** to **recognize what** is **being** said based on **visual information**
- It plays a **crucial role** in human **communication** and **speech understanding** [McGurk and MacDonald, 1976]

Lip reading: what is it and what role does it play?

- The **ability** to **recognize what** is **being** said based on **visual information**
- It plays a **crucial role** in human **communication** and **speech understanding** [McGurk and MacDonald, 1976]
  - babies selectively observe their interlocutor's vocal during social interactions [Lewkowicz and Hansen-Tift, 2012]

Lip reading: what is it and what role does it play?

- The **ability** to **recognize what** is **being** said based on **visual information**
- It plays a **crucial role** in human **communication** and **speech understanding** [McGurk and MacDonald, 1976]
  - babies selectively observe their interlocutor's vocal during social interactions [Lewkowicz and Hansen-Tift, 2012]
- It's a **difficult task** for **humans**, specially in the absence of context

Lip reading: what is it and what role does it play?

- The **ability** to **recognize what** is **being** said based on **visual information**
- It plays a **crucial role** in human **communication** and **speech understanding** [McGurk and MacDonald, 1976]
  - babies selectively observe their interlocutor's vocal during social interactions [Lewkowicz and Hansen-Tift, 2012]
- It's a **difficult task** for **humans**, specially in the absence of context
- **Multiple sounds** (phonemes) have almost **identical** lip shapes (i.e., **viseme**)

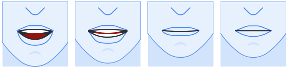


Figure 1: Bark pronunciation

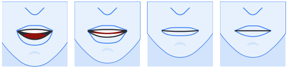


Figure 2: Mark pronunciation

Human lipreading performance is normally poor

- **Hearing-impaired** people's **accuracy** is only [Easton and Basala, 1982]

Human lipreading performance is normally poor

- **Hearing-impaired** people's **accuracy** is only [Easton and Basala, 1982]
  - $17\,\% \pm 12\,\%$ for **30 monosyllabic words**

Human lipreading performance is normally poor

- **Hearing-impaired** people's **accuracy** is only [Easton and Basala, 1982]
  - $17\,\% \pm 12\,\%$ for **30 monosyllabic words**
  - $21\,\% \pm 11\,\%$ for **30 compound words**

Human lipreading performance is normally poor

- **Hearing-impaired** people's **accuracy** is only [Easton and Basala, 1982]
  - $17\,\% \pm 12\,\%$ for **30 monosyllabic words**
  - $21\,\% \pm 11\,\%$ for **30 compound words**
- **Enormous applications** including

Human lipreading performance is normally poor

- **Hearing-impaired** people's **accuracy** is only [Easton and Basala, 1982]
  - $17\,\% \pm 12\,\%$ for **30 monosyllabic words**
  - $21\,\% \pm 11\,\%$ for **30 compound words**
- **Enormous applications** including
  - improve hearing aids

Human lipreading performance is normally poor

- **Hearing-impaired** people's **accuracy** is only [Easton and Basala, 1982]
  - $17\,\% \pm 12\,\%$ for **30 monosyllabic words**
  - $21\,\% \pm 11\,\%$ for **30 compound words**
- **Enormous applications** including
  - improve hearing aids
  - silent dictation in public spaces

Human lipreading performance is normally poor

- **Hearing-impaired** people's **accuracy** is only [Easton and Basala, 1982]
  - $17\,\% \pm 12\,\%$ for **30 monosyllabic words**
  - $21\,\% \pm 11\,\%$ for **30 compound words**
- **Enormous applications** including
  - improve hearing aids
  - silent dictation in public spaces
  - speech recognition in noisy environments

Human lipreading performance is normally poor

- **Hearing-impaired** people's **accuracy** is only [Easton and Basala, 1982]
  - $17\,\% \pm 12\,\%$ for **30 monosyllabic words**
  - $21\,\% \pm 11\,\%$ for **30 compound words**
- **Enormous applications** including
  - improve hearing aids
  - silent dictation in public spaces
  - speech recognition in noisy environments
  - salience movie processing

Human lipreading performance is normally poor

- **Hearing-impaired** people's **accuracy** is only [Easton and Basala, 1982]
  - $17\,\% \pm 12\,\%$ for **30 monosyllabic words**
  - $21\,\% \pm 11\,\%$ for **30 compound words**
- **Enormous applications** including
  - improve hearing aids
  - silent dictation in public spaces
  - speech recognition in noisy environments
  - salience movie processing
- **Automate** lipreading comprises an important **goal**

Human lipreading performance is normally poor

- **Hearing-impaired** people's **accuracy** is only [Easton and Basala, 1982]
  - $17\,\% \pm 12\,\%$ for **30 monosyllabic words**
  - $21\,\% \pm 11\,\%$ for **30 compound words**
- **Enormous applications** including
  - improve hearing aids
  - silent dictation in public spaces
  - speech recognition in noisy environments
  - salience movie processing
- **Automate** lipreading comprises an important **goal**
- **Machine lipreading** requires extracting **spatiotemporal features** from the videos

Human lipreading performance is normally poor

- **Hearing-impaired** people's **accuracy** is only [Easton and Basala, 1982]
  - $17\,\% \pm 12\,\%$ for **30 monosyllabic words**
  - $21\,\% \pm 11\,\%$ for **30 compound words**
- **Enormous applications** including
  - improve hearing aids
  - silent dictation in public spaces
  - speech recognition in noisy environments
  - salience movie processing
- **Automate** lipreading comprises an important **goal**
- **Machine lipreading** requires extracting **spatiotemporal features** from the videos
- **Deep learning** approaches offer an **end-to-end** strategy to extract these features

Outline

Speakers generalization and motion extractions were the main issues

Task

Given a silence video of a talking face, predict the sentences being spoken

Speakers generalization and motion extractions were the main issues

Task

Given a silence video of a talking face, predict the sentences being spoken

- Many works focused on video and image preprocessing to extract different features [Zhou et al., 2014]
    - Hidden Markov model (HMM) and generalized method of moments (GMM) combined with handed-engineered features
    - Speaker-dependency accuracy and/or limited utterances

Speakers generalization and motion extractions were the main issues

Task

Given a silence video of a talking face, predict the sentences being spoken

- Many works focused on video and image preprocessing to extract different features [Zhou et al., 2014]
  - Hidden Markov model (HMM) and generalized method of moments (GMM) combined with handed-engineered features
  - Speaker-dependency accuracy and/or limited utterances
- First deep learning attempts limited to word or phoneme classification
  - Fixed sequences size
  - Speaker-dependent
  - Lacked sequence prediction

Speakers generalization and motion extractions were the main issues

Task

Given a silence video of a talking face, predict the sentences being spoken

- Many works focused on video and image preprocessing to extract different features [Zhou et al., 2014]
  - Hidden Markov model (HMM) and generalized method of moments (GMM) combined with handed-engineered features
  - Speaker-dependency accuracy and/or limited utterances
- First deep learning attempts limited to word or phoneme classification
  - Fixed sequences size
  - Speaker-dependent
  - Lacked sequence prediction
- Connectionist temporal classification (CTC) loss [Graves et al., 2006]

First to show an end-to-end strategy for lipreading

- Maps variable-length **sequences** of **video** frames to **text** sequences
- GRID corpus 33k sentences

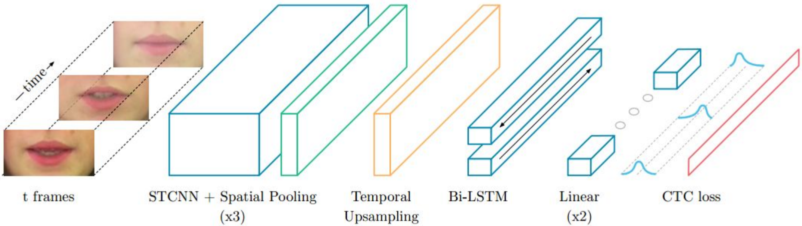


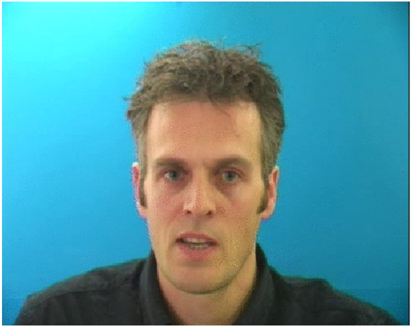
Figure 3: LipNet architecture. Source: Assael et al., 2016

GRID dataset has a fixed grammar structure

Table 1: GRID sentence and grammar structure

command	color*	preposition	letter*	digit	adverb*
{bin, lay, place, set}	{blue, green, red, white}	{at, by, in, with}	[A–Z] \ {W}	[0–9]	{again, now, please, soon}

* keywords



Four different strategies to compare with the LipNet performance

- **Hearing-impaired students** three members of the Oxford Students' Disability community
- **Baseline-LSTM**: replicate a state-of-the art architecture
- **Baseline-2D**: spatial-only convolutions
- **Baseline-NoLM**: language model disabled
- Use **word error rate (WER)** and **character error rate (CER)**

LipNet outperforms human and previous state-of-the-art model

Table 2: Performance of LipNet on the GRID dataset

Method	Unseen Speakers CER	Unseen Speakers WER	Overlapped Speakers CER	Overlapped Speakers WER
Hearing-Impaired	–	47.7%	–	–
Baseline-LSTM	38.4%	52.8%	15.2%	26.3%
Baseline-2D	16.2%	26.7%	4.3%	11.6%
Baseline-NoLM	6.7%	13.6%	2.0%	5.6%
LipNet	6.4%	11.4%	1.9%	4.8%

LipNet: takeaways

- It is an **end-to-end sentence-sequence** prediction model

LipNet: takeaways

- It is an **end-to-end sentence-sequence** prediction model
  - spatiotemporal frontend + 3D and 2D convolutions + 2 x bidirectional-LSTM (BLSTM)

LipNet: takeaways

- It is an **end-to-end sentence-sequence** prediction model
  - spatiotemporal frontend + 3D and 2D convolutions + 2 x bidirectional-LSTM (BLSTM)
- It relies on CTC to:

LipNet: takeaways

- It is an **end-to-end sentence-sequence** prediction model
  - spatiotemporal frontend + 3D and 2D convolutions + 2 x bidirectional-LSTM (BLSTM)
- It relies on CTC to:
  1. predict frame-wise labels

LipNet: takeaways

- It is an **end-to-end sentence-sequence** prediction model
  - spatiotemporal frontend + 3D and 2D convolutions + 2 x bidirectional-LSTM (BLSTM)
- It relies on CTC to:
  1. predict frame-wise labels
  2. look for the optimal alignment between the frame-wise predictions and the output sequence

LipNet: takeaways

- It is an **end-to-end sentence-sequence** prediction model
  - spatiotemporal frontend + 3D and 2D convolutions + 2 x bidirectional-LSTM (BLSTM)
- It relies on CTC to:
  1. predict frame-wise labels
  2. look for the optimal alignment between the frame-wise predictions and the output sequence
- Confirms the **importance** of combining **STCNNs** with **RNNs**

LipNet: takeaways

- It is an **end-to-end sentence-sequence** prediction model
  - spatiotemporal frontend + 3D and 2D convolutions + 2 x bidirectional-LSTM (BLSTM)
- It relies on CTC to:
  1. predict frame-wise labels
  2. look for the optimal alignment between the frame-wise predictions and the output sequence
- Confirms the **importance** of combining **STCNNs** with **RNNs**
- Extracting **spatiotemporal features** using **STCNN** is **better** than **aggregating spatial-only** features

LipNet: takeaways

- It is an **end-to-end sentence-sequence** prediction model
  - spatiotemporal frontend + 3D and 2D convolutions + 2 x bidirectional-LSTM (BLSTM)
- It relies on CTC to:
  1. predict frame-wise labels
  2. look for the optimal alignment between the frame-wise predictions and the output sequence
- Confirms the **importance** of combining **STCNNs** with **RNNs**
- Extracting **spatiotemporal features** using **STCNN** is **better** than **aggregating spatial-only** features
- **GRID dataset**: fixed grammar structure

Outline

Focus on analyzing the performance of different DL architectures

Goal

- Compare the performance and training time of three different deep learning architectures

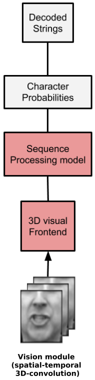


Figure 4: Deep lipreading models. Source: Afouras, Chung, and Zisserman, 2018

Focus on analyzing the performance of different DL architectures

Goal

- Compare the performance and training time of three different deep learning architectures

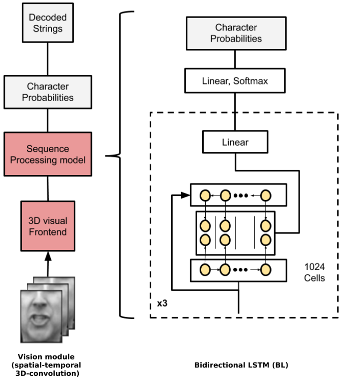


Figure 4: Deep lipreading models. Source: Afouras, Chung, and Zisserman, 2018

Focus on analyzing the performance of different DL architectures

Goal

- Compare the performance and training time of three different deep learning architectures

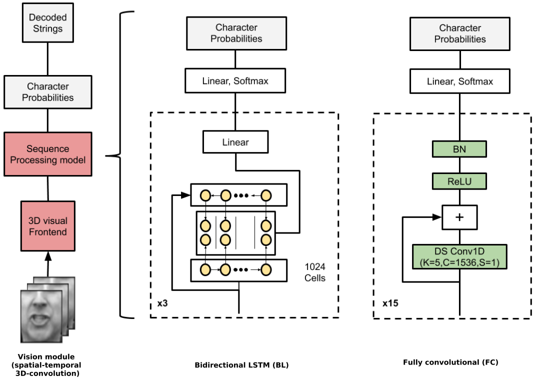


Figure 4: Deep lipreading models. Source: Afouras, Chung, and Zisserman, 2018

Focus on analyzing the performance of different DL architectures

Goal

- Compare the performance and training time of three different deep learning architectures

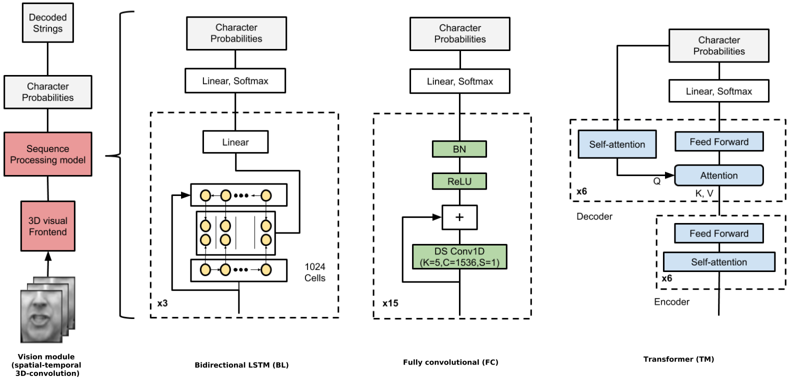
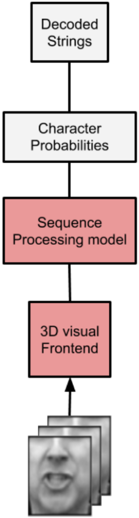


Figure 4: Deep lipreading models. Source: Afouras, Chung, and Zisserman, 2018

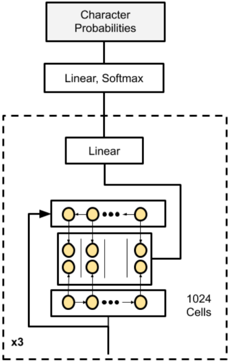
Spatiotemporal visual front-end



Vision module (spatial-temporal 3D-convolution)

- Spatiotemporal 3D convolutional on the input with a filter width of five frames
- Followed by a 2D ResNet which decreases the spatial dimensions
- For an input sequence of $T \times H \times W$ frames outputs a $T \times \frac{H}{32} \times \frac{W}{32} \times 512$ tensor
- Results in a 512-dimensional feature vector for each input video frame

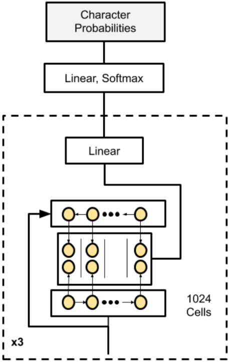
Bidirectional LSTM (BLSTM)



Bidirectional LSTM (BL)

- Comprises three stacked bidirectional LSTMs

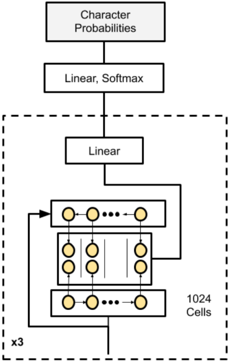
Bidirectional LSTM (BLSTM)



Bidirectional LSTM (BL)

- Comprises three stacked bidirectional LSTMs
- Ingests the video feature vectors

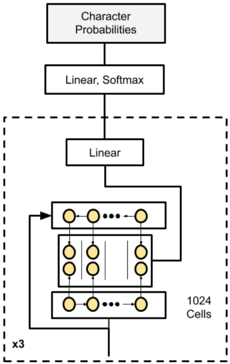
Bidirectional LSTM (BLSTM)



Bidirectional LSTM (BL)

- Comprises three stacked bidirectional LSTMs
- Ingests the video feature vectors
- Outputs a character probability for each input frame

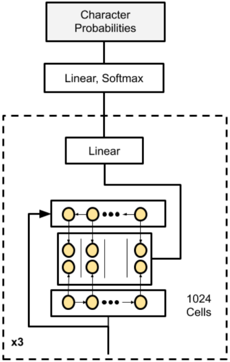
Bidirectional LSTM (BLSTM)



Bidirectional LSTM (BL)

- Comprises three stacked bidirectional LSTMs
- Ingests the video feature vectors
- Outputs a character probability for each input frame
- It's trained with connectionist temporal classification (CTC)

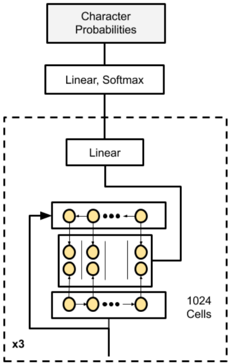
Bidirectional LSTM (BLSTM)



Bidirectional LSTM (BL)

- Comprises three stacked bidirectional LSTMs
- Ingests the video feature vectors
- Outputs a character probability for each input frame
- It's trained with connectionist temporal classification (CTC)
- Output alphabet is augmented with CTC blank character

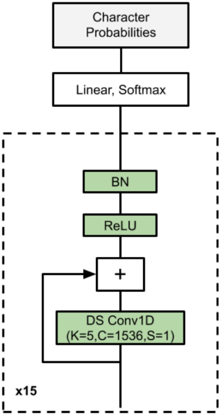
Bidirectional LSTM (BLSTM)



Bidirectional LSTM (BL)

- Comprises three stacked bidirectional LSTMs
- Ingests the video feature vectors
- Outputs a character probability for each input frame
- It's trained with connectionist temporal classification (CTC)
- Output alphabet is augmented with CTC blank character
- Decoding is performed with a beam search

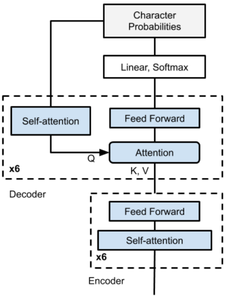
Fully convolutional (FC) model



Fully convolutional (FC)

- Rely on a depth-wise separable convolution layers
- Each convolution adds a skip-connection followed by ReLU and batch normalization
- Also trained with CTC loss
- Considers two variants: 10 and 15 convolutional layers

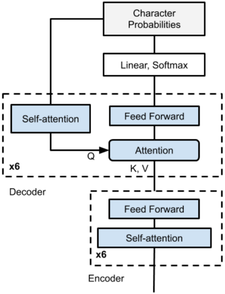
Transformer model (TC)



Transformer (TM)

- Input serves as attention queries, keys, and values

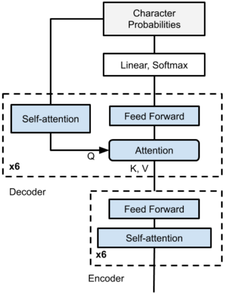
Transformer model (TC)



Transformer (TM)

- Input serves as attention queries, keys, and values
- Encoder outputs are the the attention keys and values

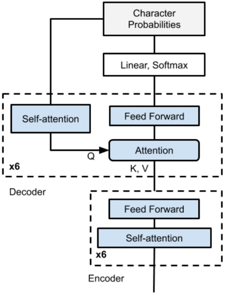
Transformer model (TC)



Transformer (TM)

- Input serves as attention queries, keys, and values
- Encoder outputs are the the attention keys and values
- Previous decoding layer outputs are the queries

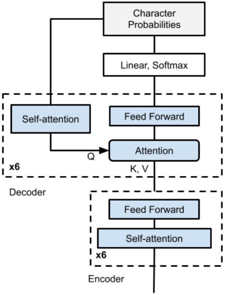
Transformer model (TC)



Transformer (TM)

- Input serves as attention queries, keys, and values
- Encoder outputs are the the attention keys and values
- Previous decoding layer outputs are the queries
- The decoder produces character probabilities

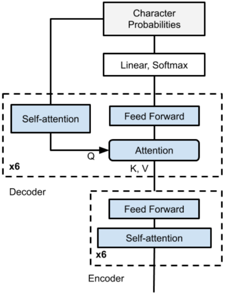
Transformer model (TC)



Transformer (TM)

- Input serves as attention queries, keys, and values
- Encoder outputs are the the attention keys and values
- Previous decoding layer outputs are the queries
- The decoder produces character probabilities
- Rely on the based model proposed by Vaswani et al., 2017

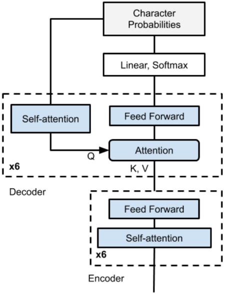
Transformer model (TC)



Transformer (TM)

- Input serves as attention queries, keys, and values
- Encoder outputs are the the attention keys and values
- Previous decoding layer outputs are the queries
- The decoder produces character probabilities
- Rely on the based model proposed by Vaswani et al., 2017
  - 6 encoder and 6 decoder layers

Transformer model (TC)



Transformer (TM)

- Input serves as attention queries, keys, and values
- Encoder outputs are the the attention keys and values
- Previous decoding layer outputs are the queries
- The decoder produces character probabilities
- Rely on the based model proposed by Vaswani et al., 2017
  - 6 encoder and 6 decoder layers
  - 8 attention heads with dropout with $p = 0.1$

An external character-level language model

- Use a character-level language model during inference
- Recurrent neural network with 4 unidirectional layers of 1024 LSTM cells each
- Trained to predict one character at a time

Two different datasets for performance evaluation

Table 3: Datasets used for trained and test

Dataset	# Words	Type	Vocabulary	# Utter.	Viewpoint
LRW	489k	single word	500	–	unique
LRS2	2M	sentences	41K	142K	multiple
MV-LRS(w)	1.9M	sentences	480	–	unique
MV-LRS	5M	sentences	30K	430K	unique

LRW: Lip Reading in the Wild
LRS2: Lip Reading Sentences 2

- Two different corpora to train the language model

Two different datasets for performance evaluation

Table 3: Datasets used for trained and test

Dataset	# Words	Type	Vocabulary	# Utter.	Viewpoint
LRW	489k	single word	500	–	unique
LRS2	2M	sentences	41K	142K	multiple
MV-LRS(w)	1.9M	sentences	480	–	unique
MV-LRS	5M	sentences	30K	430K	unique

LRW: Lip Reading in the Wild
LRS2: Lip Reading Sentences 2

- Two different corpora to train the language model
  1. transcriptions of the LRS2 pre-train and main train-data $\equiv 2M$ words

Two different datasets for performance evaluation

Table 3: Datasets used for trained and test

Dataset	# Words	Type	Vocabulary	# Utter.	Viewpoint
LRW	489k	single word	500	–	unique
LRS2	2M	sentences	41K	142K	multiple
MV-LRS(w)	1.9M	sentences	480	–	unique
MV-LRS	5M	sentences	30K	430K	unique

LRW: Lip Reading in the Wild
LRS2: Lip Reading Sentences 2

- Two different corpora to train the language model
  1. transcriptions of the LRS2 pre-train and main train-data $\equiv 2M$ words
  2. full subtitles of LRS2 training set $\equiv 26M$ words

Two different datasets for performance evaluation

Table 3: Datasets used for trained and test

Dataset	# Words	Type	Vocabulary	# Utter.	Viewpoint
LRW	489k	single word	500	–	unique
LRS2	2M	sentences	41K	142K	multiple
MV-LRS(w)	1.9M	sentences	480	–	unique
MV-LRS	5M	sentences	30K	430K	unique

LRW: Lip Reading in the Wild
LRS2: Lip Reading Sentences 2

- Two different corpora to train the language model
  1. transcriptions of the LRS2 pre-train and main train-data $\equiv 2M$ words
  2. full subtitles of LRS2 training set $\equiv 26M$ words
- Evaluated on LRS2 $\equiv 1,243$ utterances

Two different datasets for performance evaluation

Table 3: Datasets used for trained and test

Dataset	# Words	Type	Vocabulary	# Utter.	Viewpoint
LRW	489k	single word	500	–	unique
LRS2	2M	sentences	41K	142K	multiple
MV-LRS(w)	1.9M	sentences	480	–	unique
MV-LRS	5M	sentences	30K	430K	unique

LRW: Lip Reading in the Wild
LRS2: Lip Reading Sentences 2

- Two different corpora to train the language model
  1. transcriptions of the LRS2 pre-train and main train-data $\equiv 2M$ words
  2. full subtitles of LRS2 training set $\equiv 26M$ words
- Evaluated on LRS2 $\equiv 1,243$ utterances
- Report **character error rates (CER)** and **word error rates (WER)**

Training process includes three stages

1. Visual front-end module
2. Use vision module to generate visual features for all the training data
3. Sequence processing module

Transformer architecture seems to be good choice

Table 4: Character error rates and word error rates on LRS2 dataset

Net	Method	# p	CER Greedy	CER T2	WER Greedy	WER T1	WER T2	t/b (s)	time
B	MV-WAS [15]	-	-	-	-	70.4%	-	-	-
BL	BLSTM + CTC	67M	40.6%	38.0%	76.5%	62.9%	62.2%	0.76	4.5d
FC-10	FC×10 + CTC	24M	37.1%	35.0%	69.1%	58.2%	57.1%	0.23	2.4d
FC-15	FC×15 + CTC	35M	35.3%	33.9%	64.8%	56.3%	55.0%	0.34	3.4d
TM	Transformer	40M	38.6%	34.0%	58.0%	51.2%	**50.0%**	0.41	13d

lower is better

- Transformer outperforms the other network models

Transformer architecture seems to be good choice

Table 4: Character error rates and word error rates on LRS2 dataset

Net	Method	# p	CER Greedy	CER T2	WER Greedy	WER T1	WER T2	t/b (s)	time
B	MV-WAS [15]	-	-	-	-	70.4%	-	-	-
BL	BLSTM + CTC	67M	40.6%	38.0%	76.5%	62.9%	62.2%	0.76	4.5d
FC-10	FC×10 + CTC	24M	37.1%	35.0%	69.1%	58.2%	57.1%	0.23	2.4d
FC-15	FC×15 + CTC	35M	35.3%	33.9%	64.8%	56.3%	55.0%	0.34	3.4d
TM	Transformer	40M	38.6%	34.0%	58.0%	51.2%	**50.0%**	0.41	13d

lower is better

- Transformer outperforms the other network models
- An improvement of 20% over previous state-of-the-art model

Transformer architecture seems to be good choice

Table 4: Character error rates and word error rates on LRS2 dataset

Net	Method	# p	CER Greedy	CER T2	WER Greedy	WER T1	WER T2	t/b (s)	time
B	MV-WAS [15]	-	-	-	-	70.4%	-	-	-
BL	BLSTM + CTC	67M	40.6%	38.0%	76.5%	62.9%	62.2%	0.76	4.5d
FC-10	FC×10 + CTC	24M	37.1%	35.0%	69.1%	58.2%	57.1%	0.23	2.4d
FC-15	FC×15 + CTC	35M	35.3%	33.9%	64.8%	56.3%	55.0%	0.34	3.4d
TM	Transformer	40M	38.6%	34.0%	58.0%	51.2%	**50.0%**	0.41	13d

lower is better

- Transformer outperforms the other network models
- An improvement of 20% over previous state-of-the-art model
- High computational cost (i.e., 13 days to train the model)

Takeaways

- Lipreading is a **challenge** problem

Takeaways

- Lipreading is a **challenge** problem
- **Context** information **plays** an important **role**

Takeaways

- Lipreading is a **challenge** problem
- **Context** information **plays** an important **role**
- **Transformer** architecture combined with **convolutional** neural networks **enable** machine lipreading

Takeaways

- Lipreading is a **challenge** problem
- **Context** information **plays** an important **role**
- **Transformer** architecture combined with **convolutional** neural networks **enable** machine lipreading
- **Machine lipreading** can **outperform** human-performance

Takeaways

- Lipreading is a **challenge** problem
- **Context** information **plays** an important **role**
- **Transformer** architecture combined with **convolutional** neural networks **enable** machine lipreading
- **Machine lipreading** can **outperform** human-performance
- **Computational cost** is still an **issue**

References

1. H. McGurk and J. MacDonald. "Hearing lips and seeing voices". In: *Nature* 264.5588 (1976), pp. 746–748
2. D. J. Lewkowicz and A. M. Hansen-Tift. "Infants deploy selective attention to the mouth of a talking face when learning speech". In: *National Academy of Sciences* 109.5 (2012), pp. 1431–1436
3. R. D. Easton and M. Basala. "Perceptual dominance during lipreading". In: *Perception & Psychophysics* 32.6 (1982), pp. 562–570
4. A. Vaswani et al. "Attention is all you need". In: *NeurIPS*. 2017, pp. 5998–6008
5. T. Afouras, J. S. Chung, and A. Zisserman. "Deep Lip Reading: a comparison of models and an online application". In: *INTERSPEECH*. 2018


That's all Folks!