Lip Reading Sentences in the Wild

Joon Son Chung, Andrew Senior, Oriol Vinyals, Andrew Zisserman
Presented by: Michael Fang

Task: Lip Reading

+

→ The cat sat...


MANCHESTER CITY TRAVEL TO PARIS IN THE CHAMPIONS LEAGUE THIS EVENING

Outline

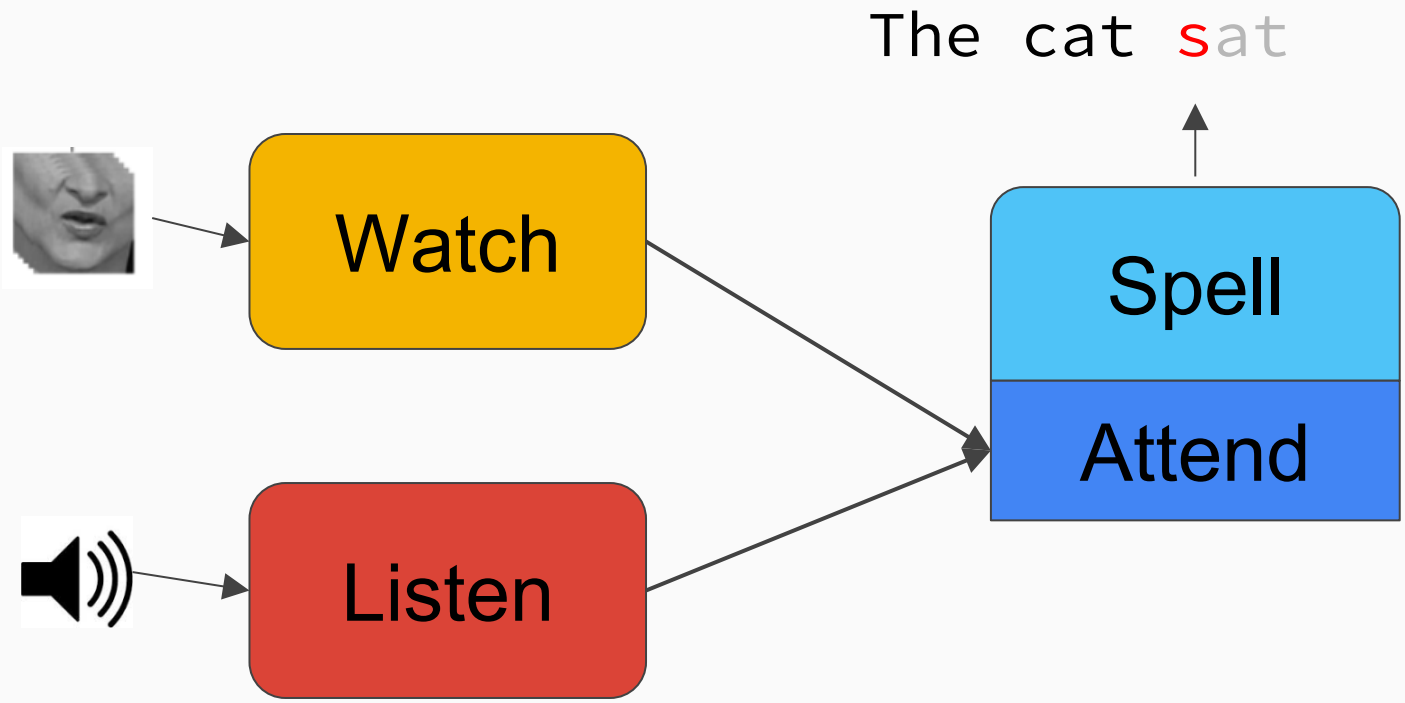
Model Architecture: Watch, Listen, Attend and Spell

Training Strategies

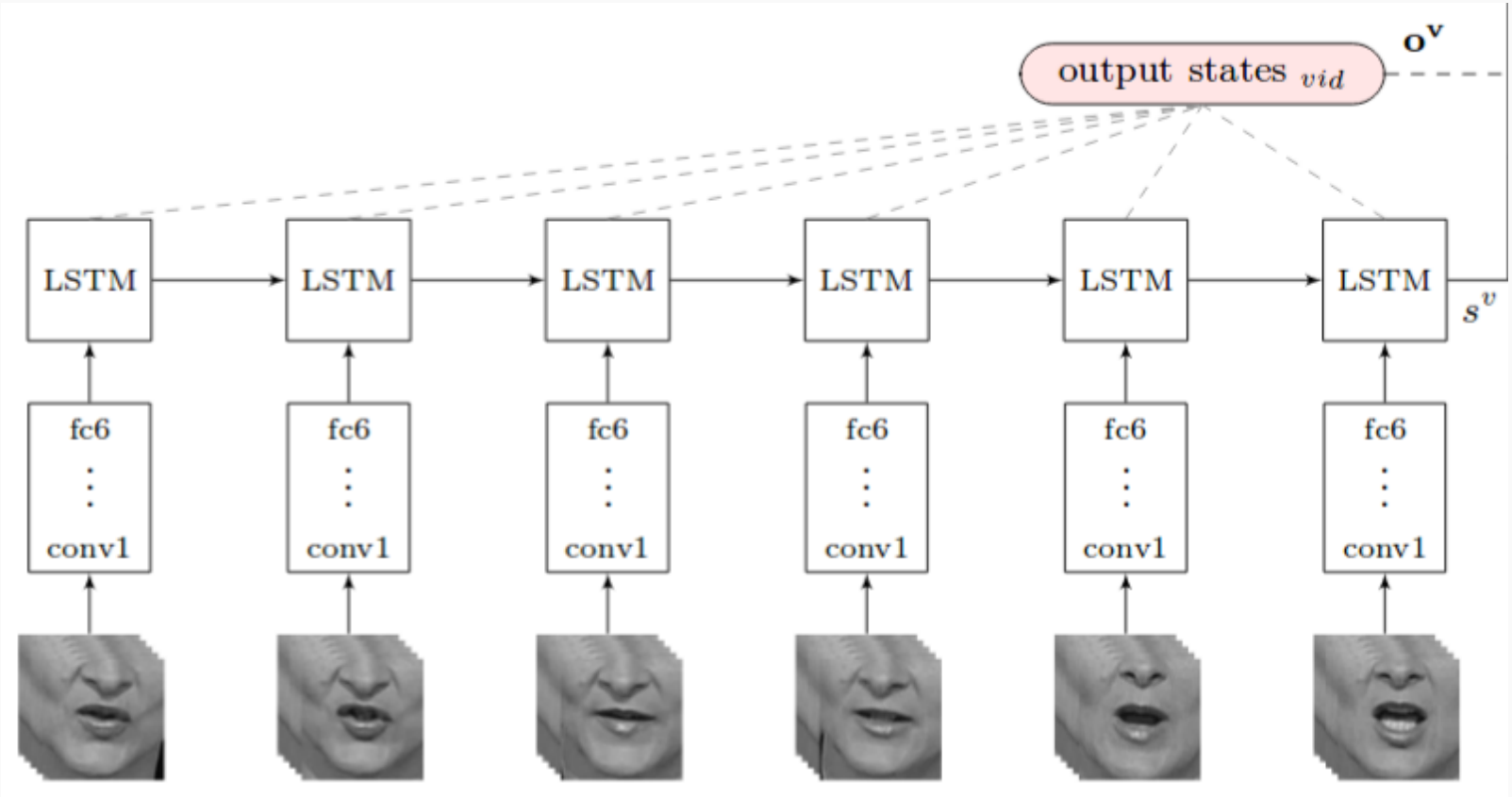
Dataset

(Professional-Surpassing!) Results

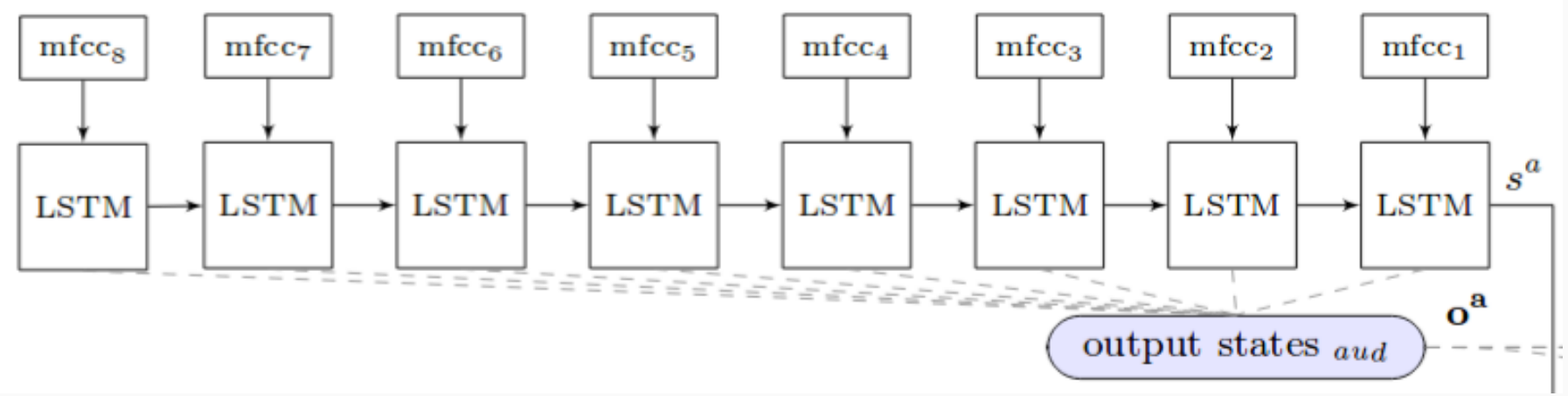
Architecture



Watch

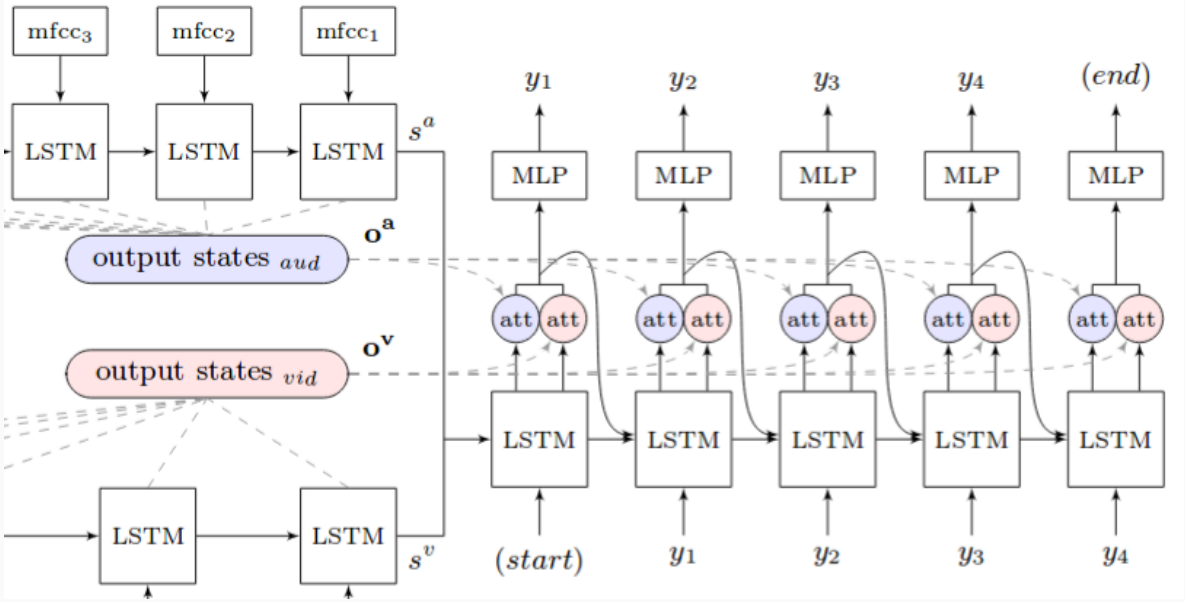

output states $_{vid}$
$\mathbf{o^v}$
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
s^v
fc6
conv1
fc6
conv1
fc6
conv1
fc6
conv1
fc6
conv1
fc6
conv1

Listen


mfcc8
mfcc7
mfcc6
mfcc5
mfcc4
mfcc3
mfcc2
mfcc1
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
s^a
$\mathbf{o^a}$
output states aud

Attend and Spell

$$h_k^d, o_k^d = \texttt{LSTM}(h_{k-1}^d, y_{k-1}, c_{k-1}^v, c_{k-1}^a)$$
$$c_k^v = \mathbf{o}^v \cdot \texttt{Attention}^{\texttt{v}}(h_k^d, \mathbf{o}^v)$$
$$c_k^a = \mathbf{o}^a \cdot \texttt{Attention}^{\texttt{a}}(h_k^d, \mathbf{o}^a)$$
$$P(y_i|\mathbf{x}^v, \mathbf{x}^a, y_{<i}) = \texttt{softmax}(\texttt{MLP}(o_k^d, c_k^v, c_k^a))$$



Curriculum Learning

Slowly increase the length of training sequences

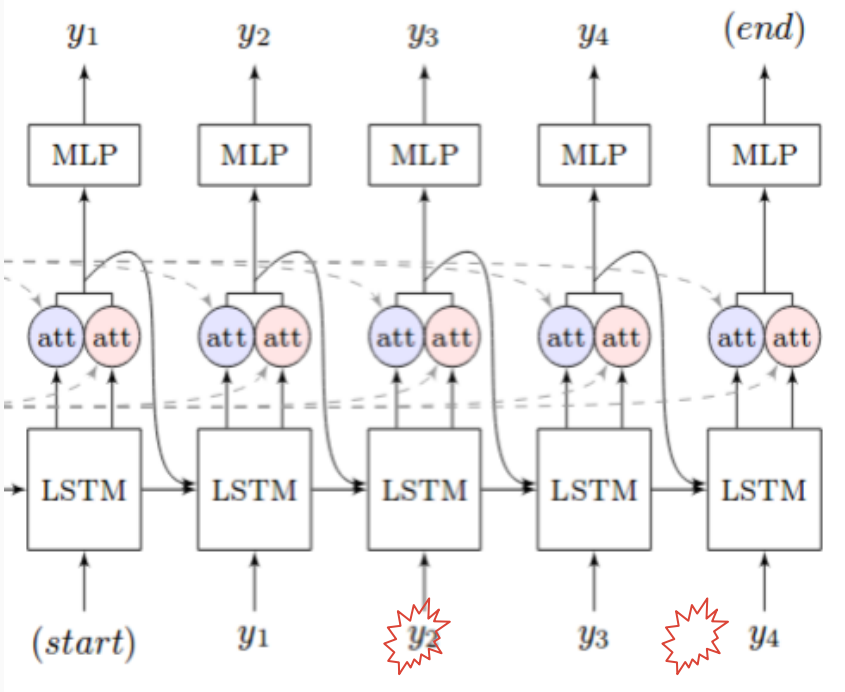
Converges training faster, decreases overfitting

```
The
The cat
The cat sat
The cat sat on
The cat sat on the
The cat sat on the mat
```

Scheduled Sampling

Randomly sample from previous prediction instead of ground truth during training

Makes training scenario more similar to testing



Dataset

Channel	Series name	# hours	# sent.
BBC 1 HD	News†	1,584	50,493
BBC 1 HD	Breakfast	1,997	29,862
BBC 1 HD	Newsnight	590	17,004
BBC 2 HD	World News	194	3,504
BBC 2 HD	Question Time	323	11,695
BBC 4 HD	World Today	272	5,558
All		4,960	118,116

Results

Method	SNR	CER	WER	BLEU†
Lips only				
Professional‡	-	58.7%	73.8%	23.8
WAS	-	59.9%	76.5%	35.6
WAS+CL	-	47.1%	61.1%	46.9
WAS+CL+SS	-	42.4%	58.1%	50.0
WAS+CL+SS+BS	-	39.5%	50.2%	54.9
Audio only				
Google Speech API	clean	17.6%	22.6%	78.4
Kaldi SGMM+MMI⋆	clean	9.7%	16.8%	83.6
LAS+CL+SS+BS	clean	10.4%	17.7%	84.0
LAS+CL+SS+BS	10dB	26.2%	37.6%	66.4
LAS+CL+SS+BS	0dB	50.3%	62.9%	44.6
Audio and lips				
WLAS+CL+SS+BS	clean	7.9%	13.9%	87.4
WLAS+CL+SS+BS	10dB	17.6%	27.6%	75.3
WLAS+CL+SS+BS	0dB	29.8%	42.0%	63.1