

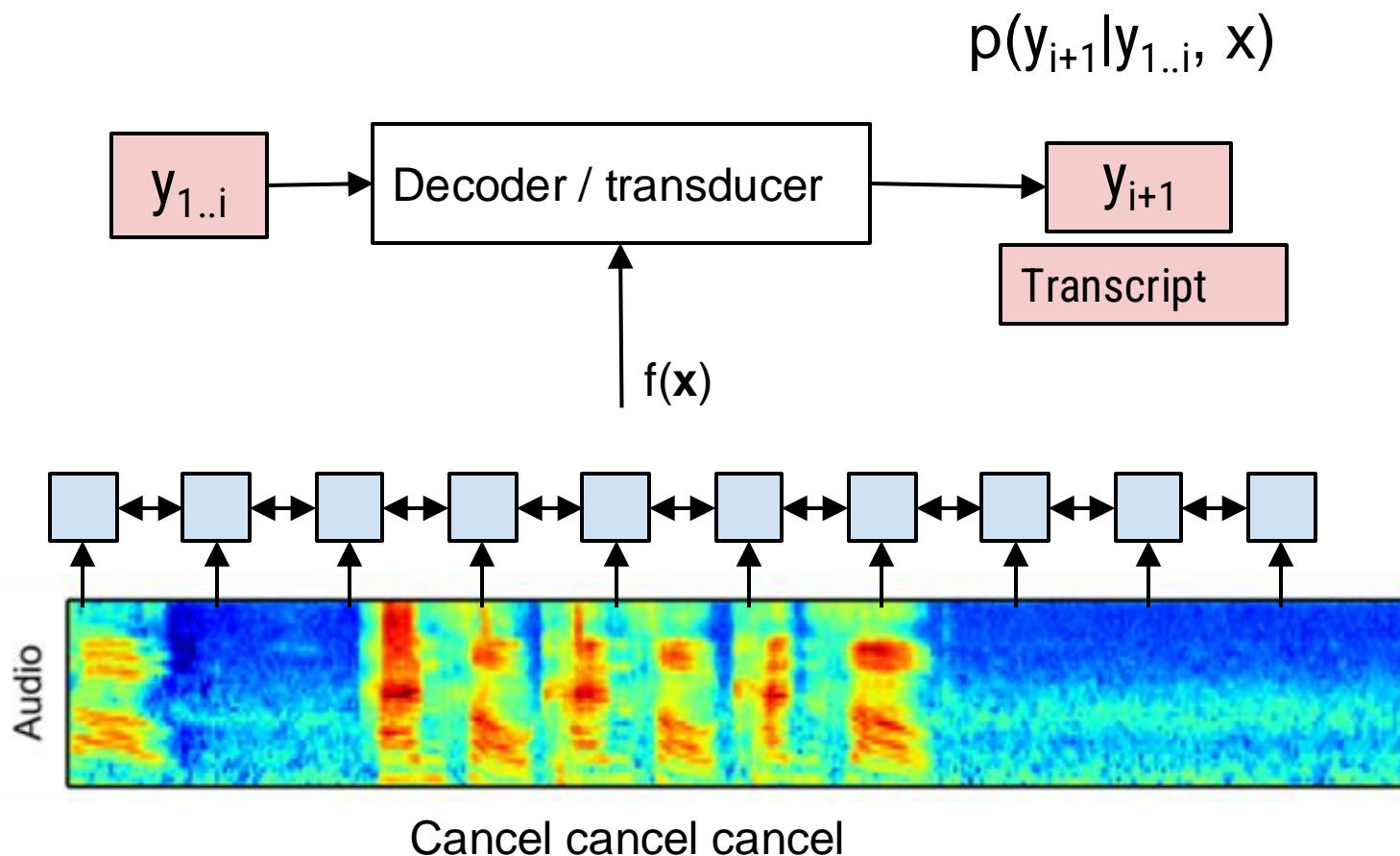
10707
Deep Learning
Russ Salakhutdinov

Machine Learning Department
rsalakhu@cs.cmu.edu

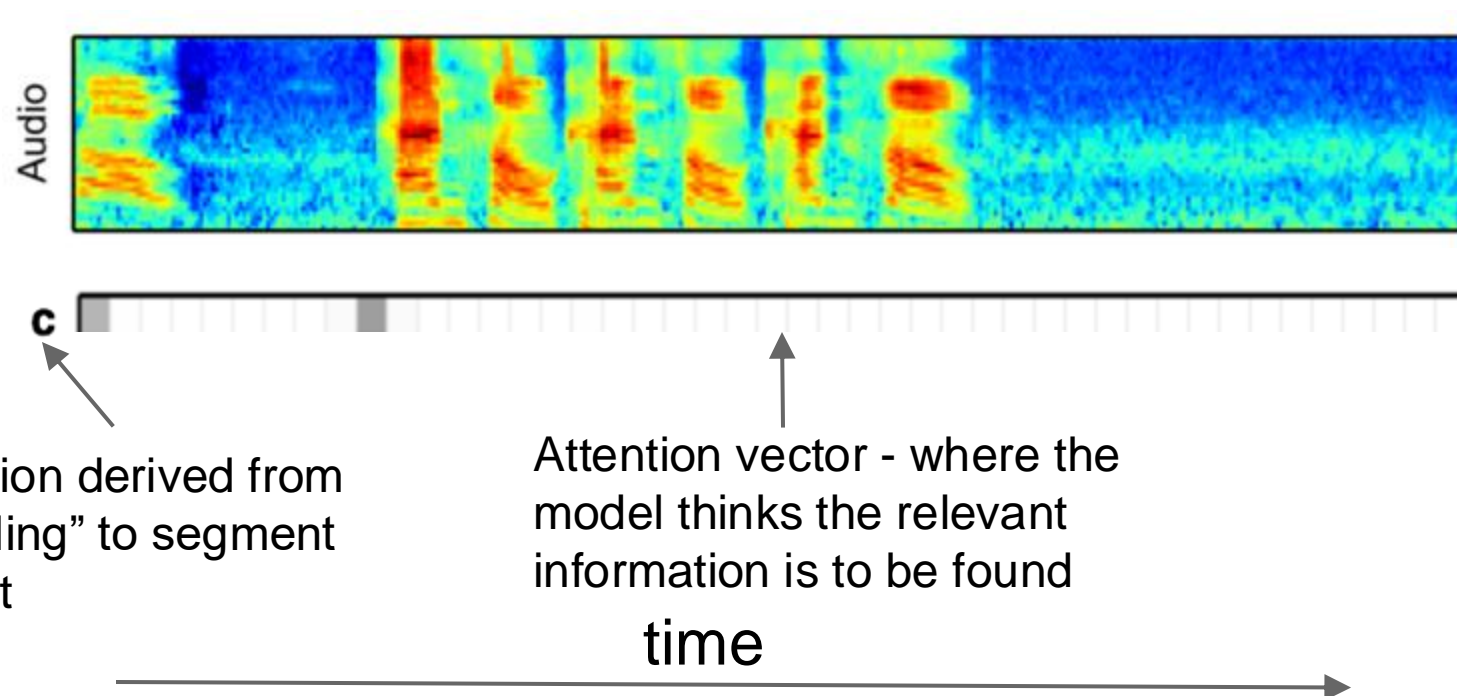
Sequence Model / Transformers

Applications

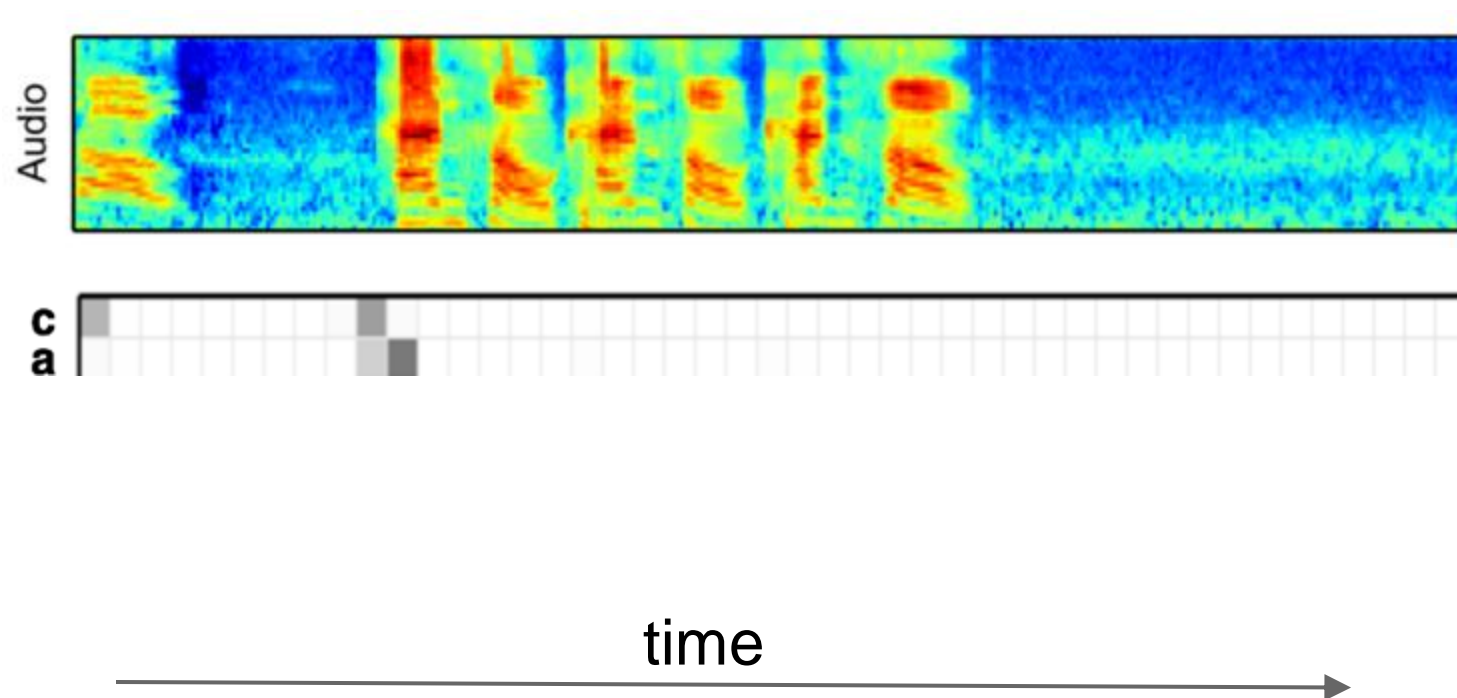
Speech Recognition



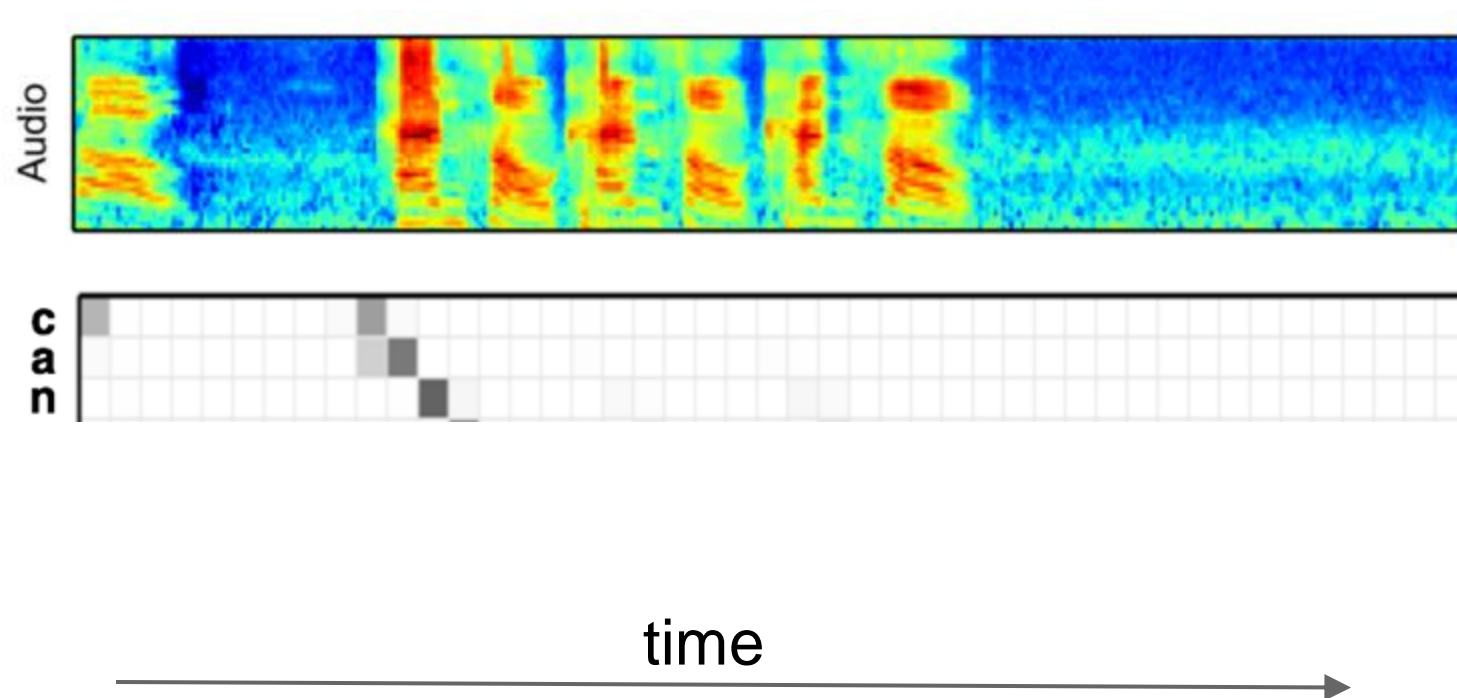
Attention Example



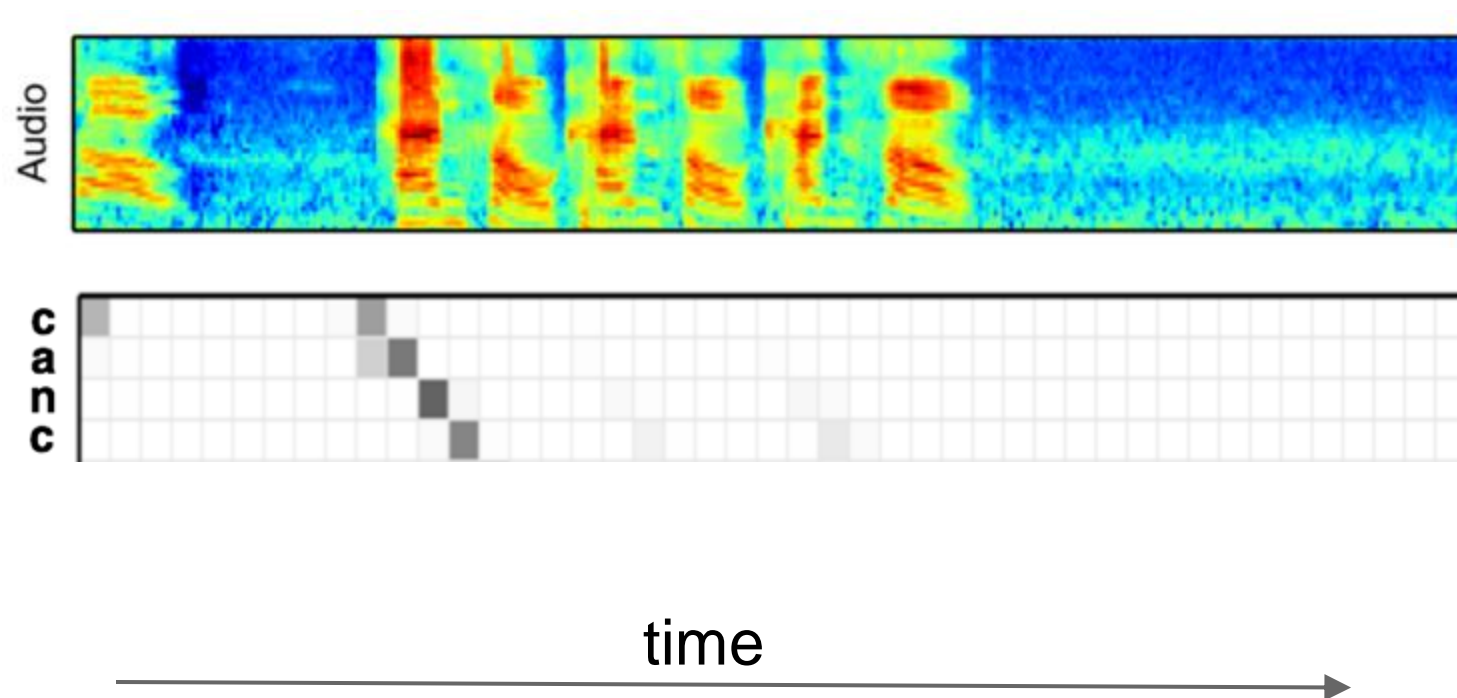
Attention Example



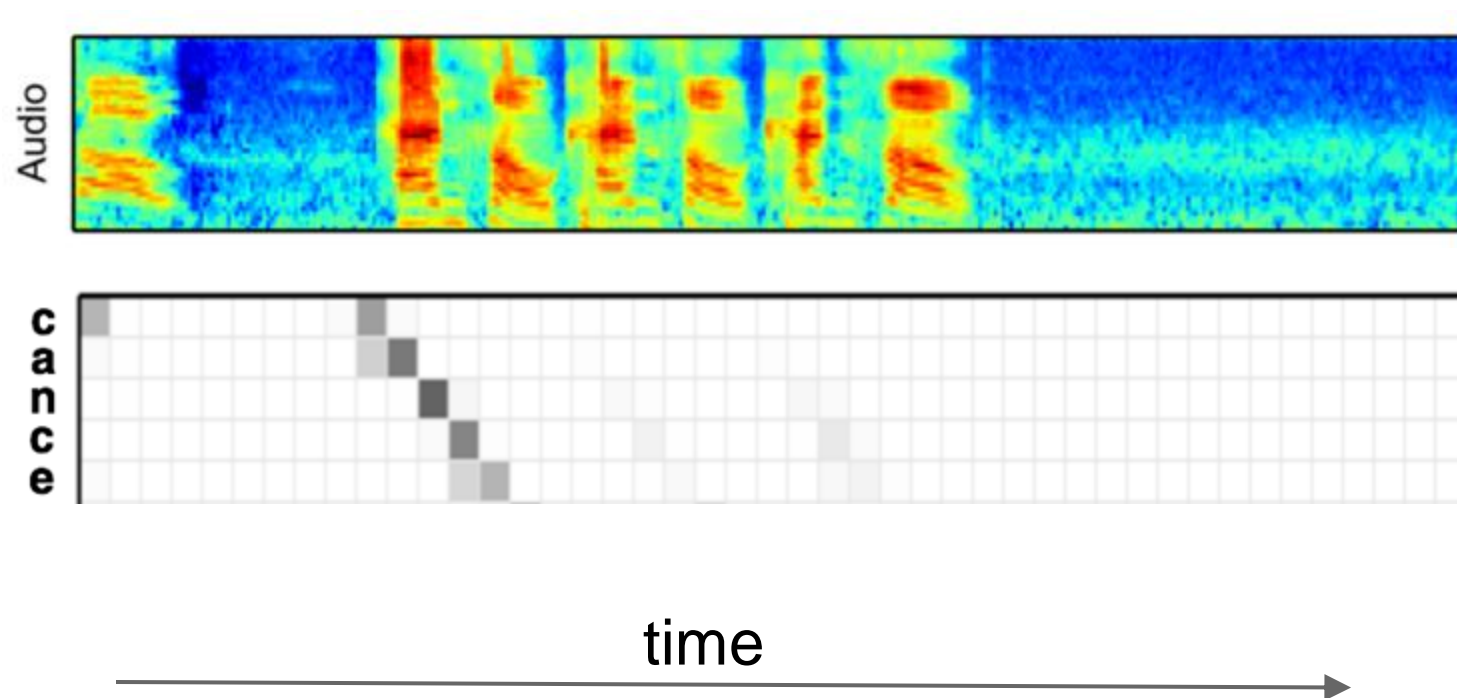
Attention Example



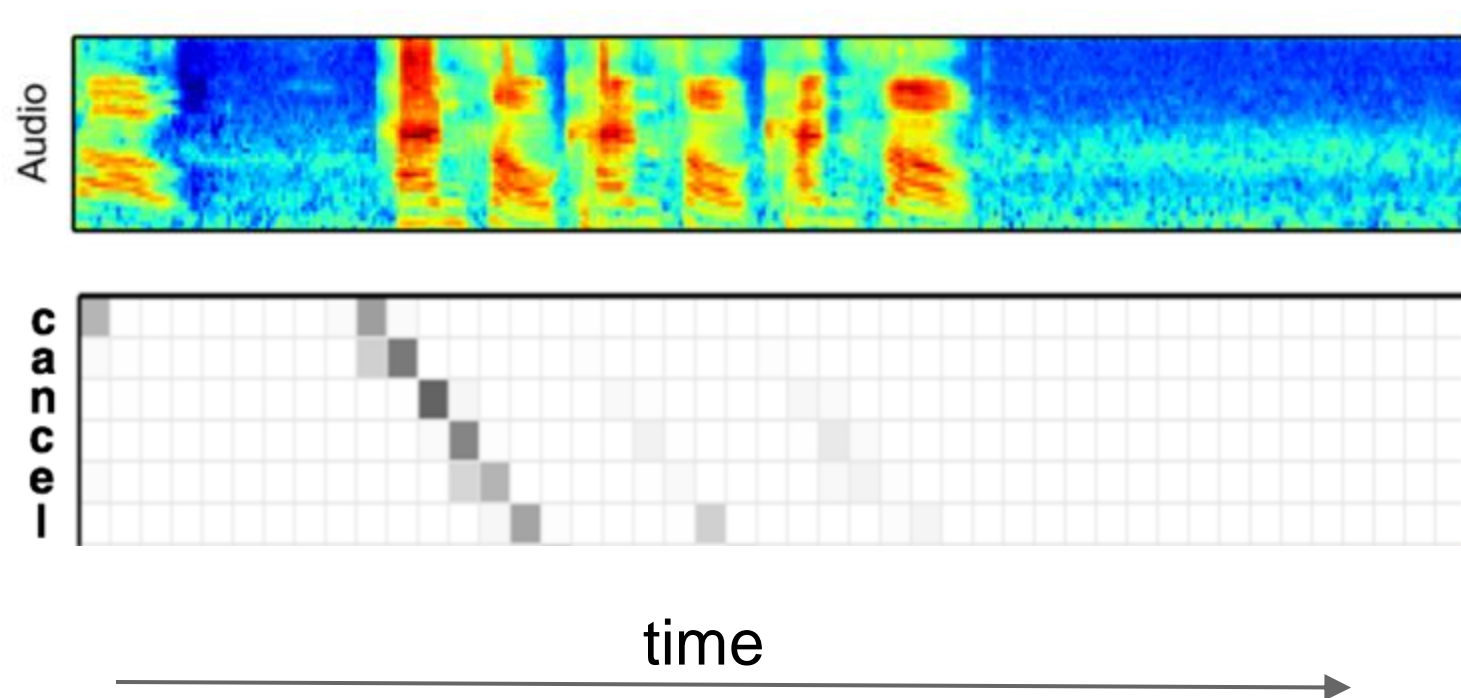
Attention Example



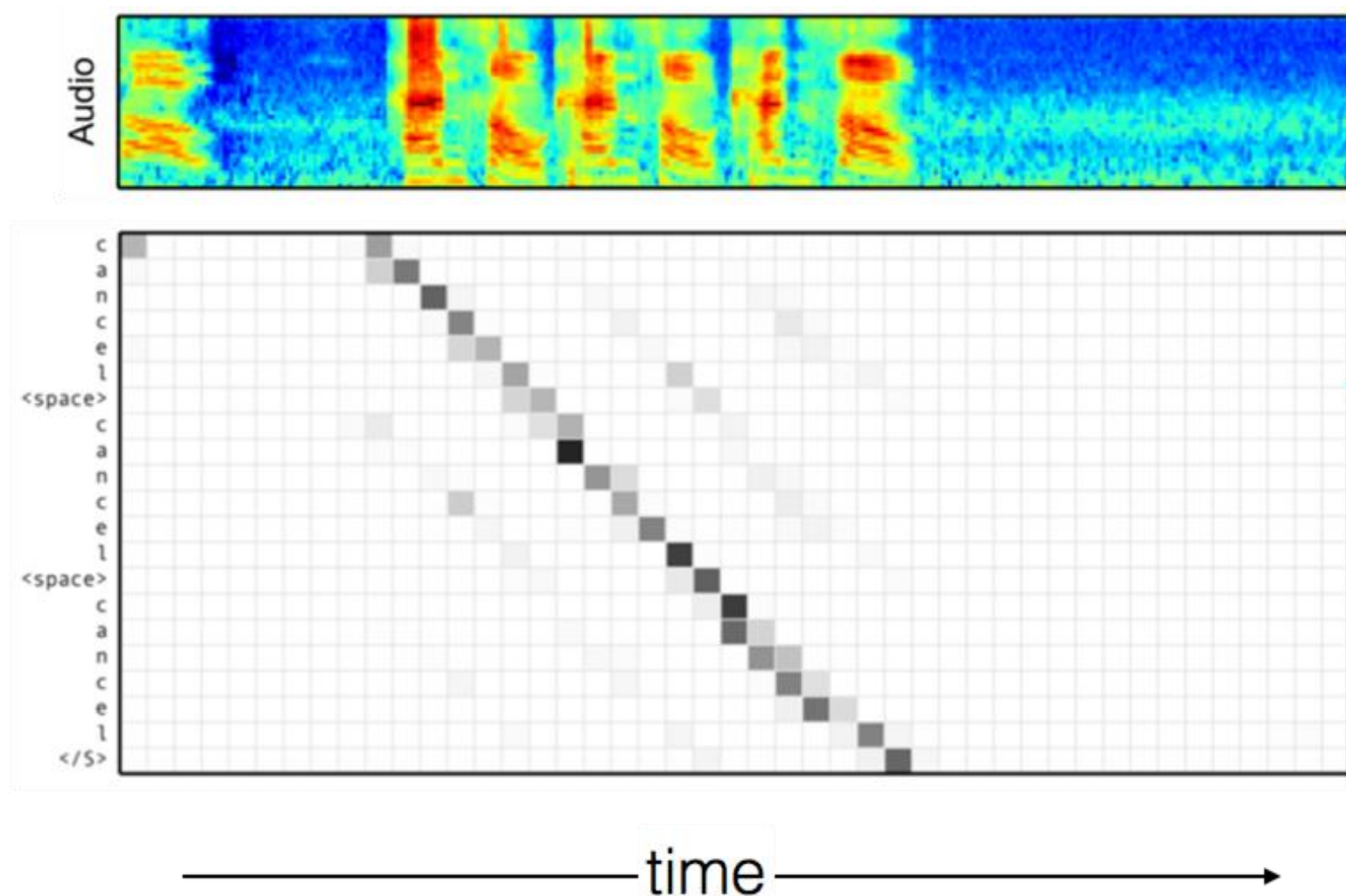
Attention Example



Attention Example



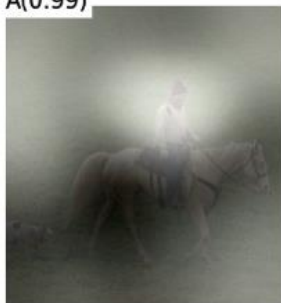
Attention Example



Caption Generation with Visual Attention



A(0.99)



man(0.40)



riding(0.26)



a(0.17)



horse(0.24)



in(0.24)



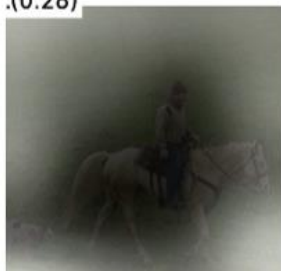
a(0.14)



field(0.43)



.(0.28)



A man riding a
horse in a field.

Caption Generation with Visual Attention



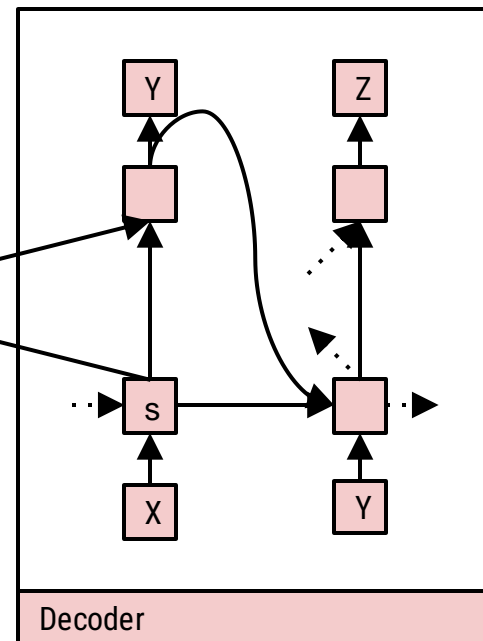
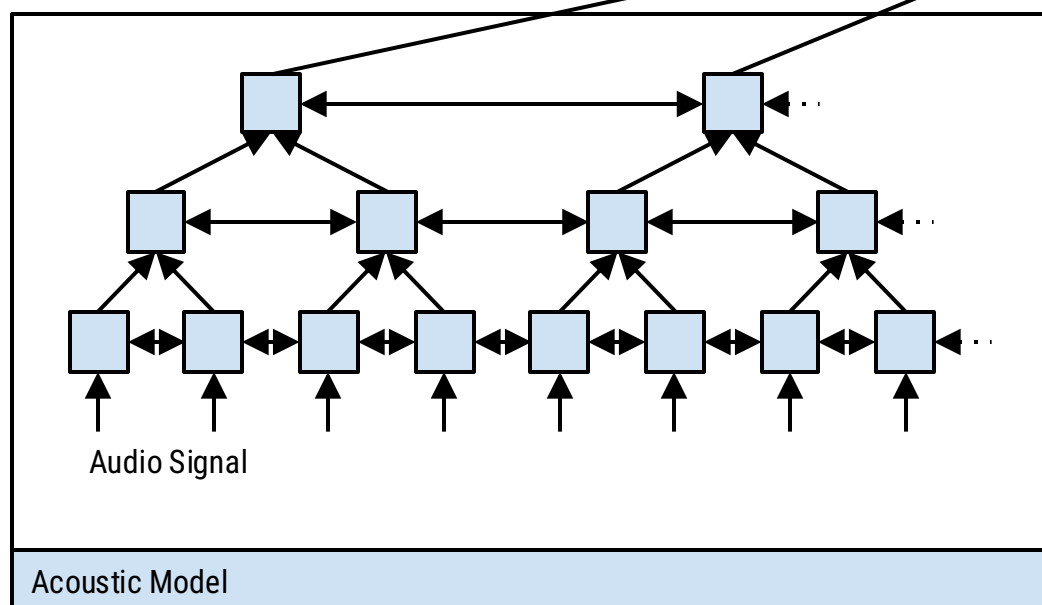
A woman holding a clock in her hand.



A large white bird standing in a forest.

Listen Attend and Spell (LAS)

- Reducing time resolution with a pyramidal encoder



LAS Results

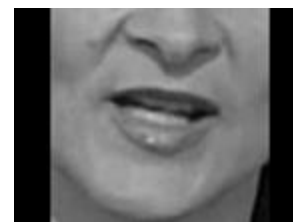
Beam	Text	LogProb	WER
Truth	call aaa roadside assistance	-	-
1	call aaa roadside assistance	-0.5740	0.00
2	call triple a roadside assistance	-1.5399	50.0
3	call trip way roadside assistance	-3.5012	50.0
4	call xxx roadside assistance	-4.4375	25.0

Lip Reading

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



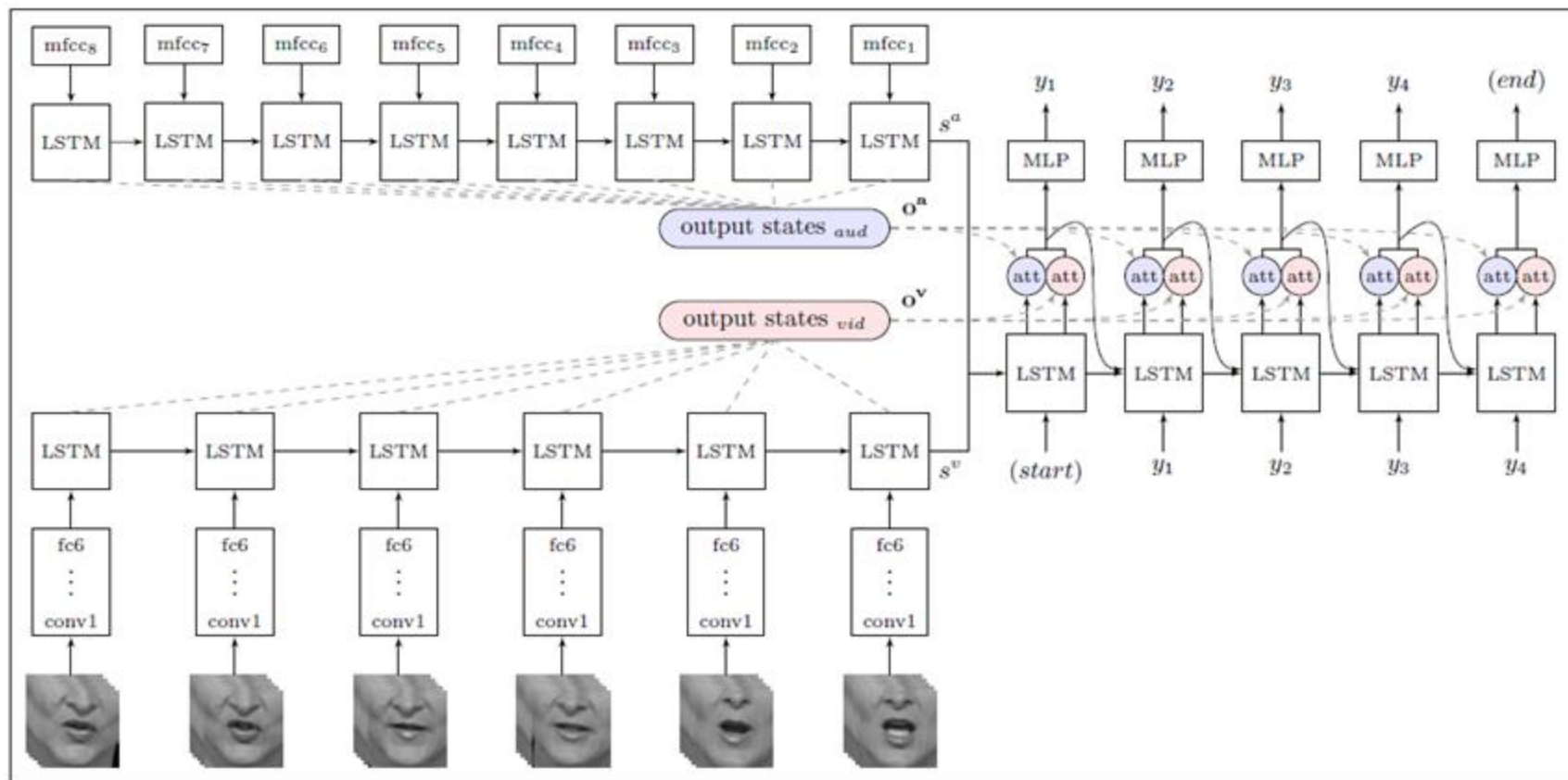
http://www.robots.ox.ac.uk/~vgg/data/lip_reading/



1. Chung, J., et al. "Lip reading sentences in the wild." *CVPR* (2017).
2. Assael, Y., et al. "Lipnet: Sentence-level lipreading." *arxiv* (2016).

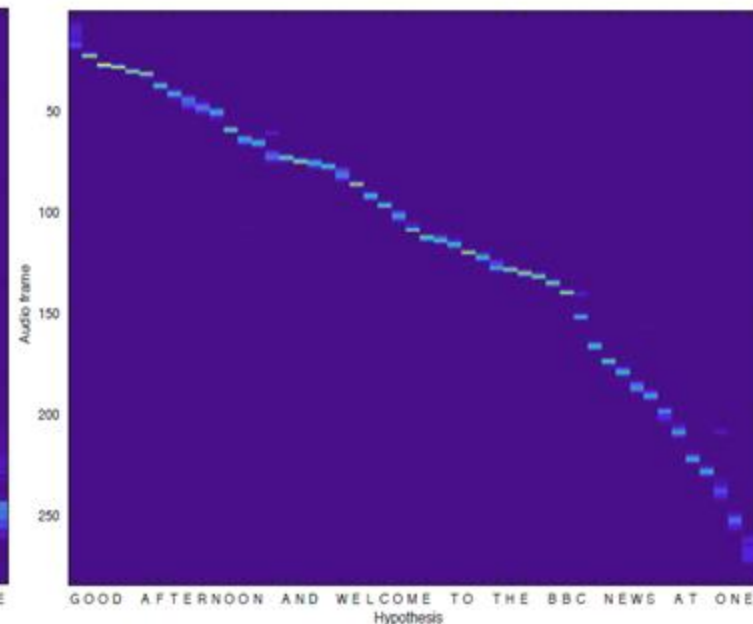
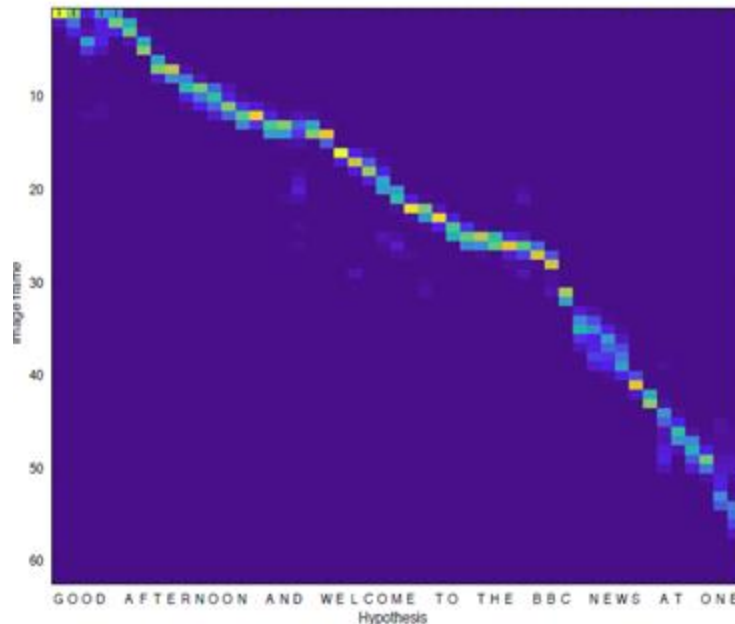
Lip Reading

Separate embedding and attention for audio and visual streams

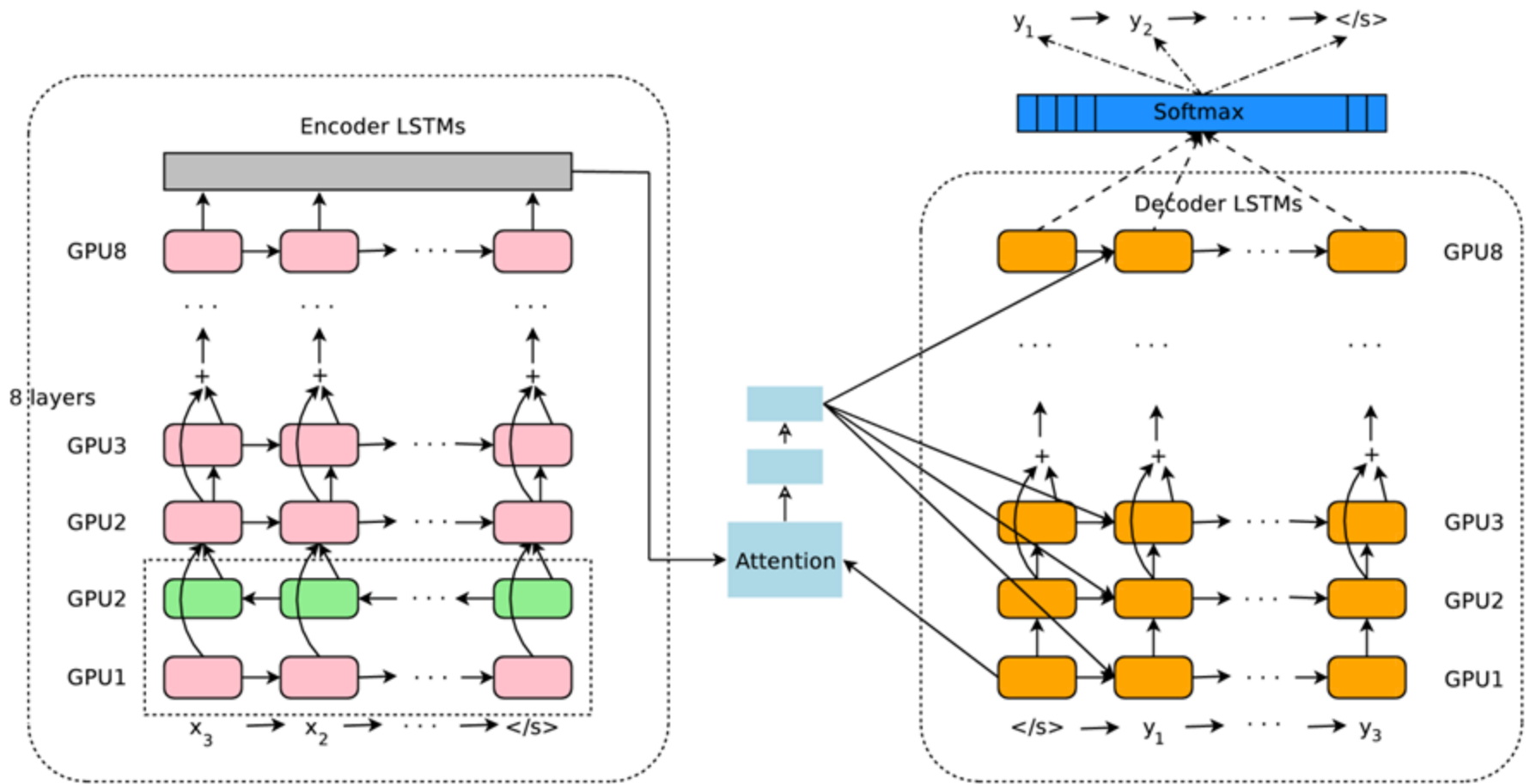


Lip Reading

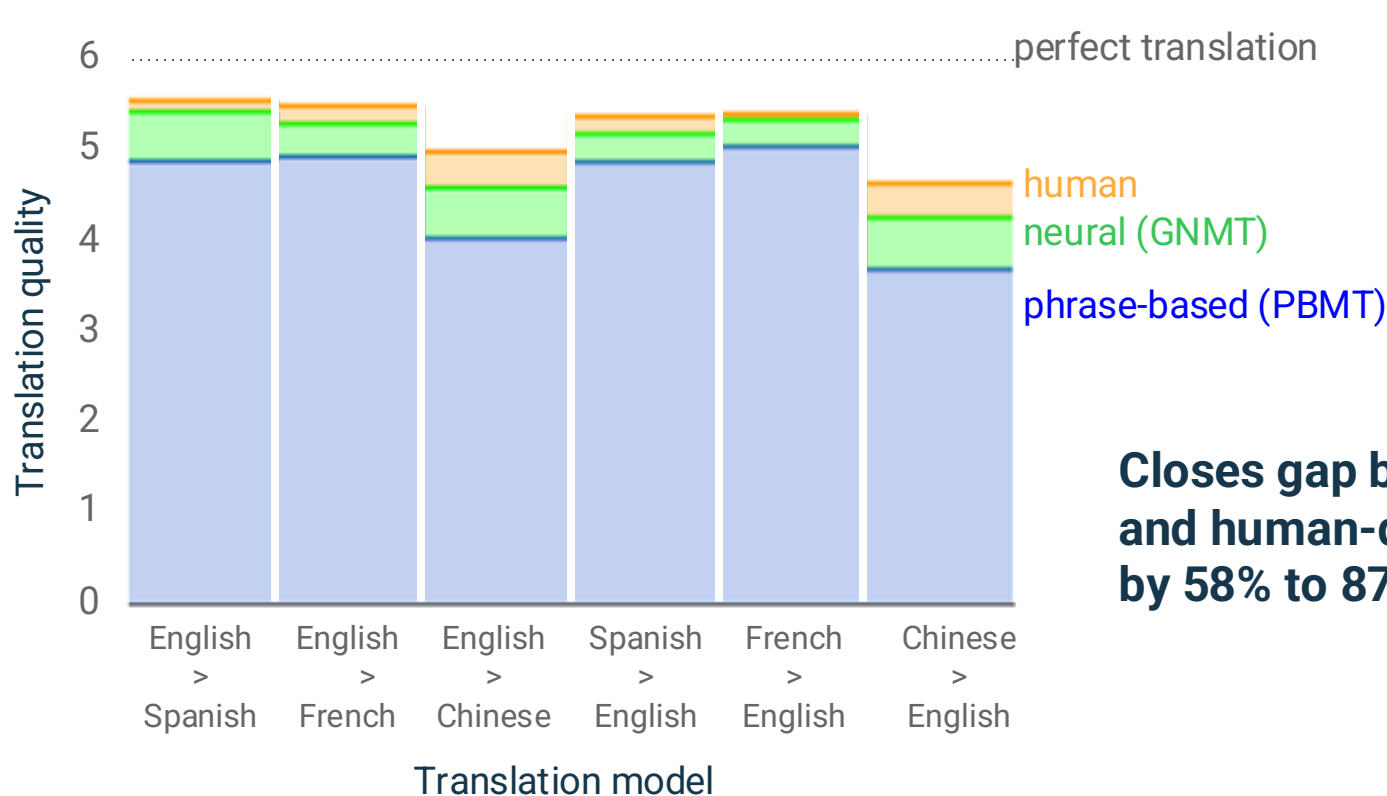
Method	SNR	CER	WER	BLEU [†]
Lips only				
Professional [‡]	-	58.7%	73.8%	23.8
WAS	-	59.9%	76.5%	35.6
WAS+CL	-	47.1%	61.1%	46.9
WAS+CL+SS	-	42.4%	58.1%	50.0
WAS+CL+SS+BS	-	39.5%	50.2%	54.9



Google Neural Machine Translation System



Google Neural Machine Translation System



Closes gap between old system and human-quality translation by 58% to 87%

Loss Functions

Loss Functions

- Cross Entropy
- Scheduled Sampling [1]
- Expected Loss [2]
- Augmented Loss [3]
- Sequence to Sequence as a beam search optimization [4]
- Learning decoders with different loss function [5]

1. Bengio, S., et al. "Scheduled sampling for sequence prediction with recurrent neural networks." *NIPS (2015)*.
2. Ranzato, M., et al. "Sequence level training with recurrent neural networks." *ICLR (2016)*.
3. Norouzi, M., et al. "Reward augmented maximum likelihood for neural structured prediction." *NIPS (2016)*.
4. Wiseman, S., Rush, A. "Sequence-to-sequence learning as beam-search optimization." *EMLP (2016)*.
5. Gu, J, Cho, K and Li, V.O.K. "Trainable greedy decoding for neural machine translation." *arXiv preprint arXiv:1702.02429 (2017)*.

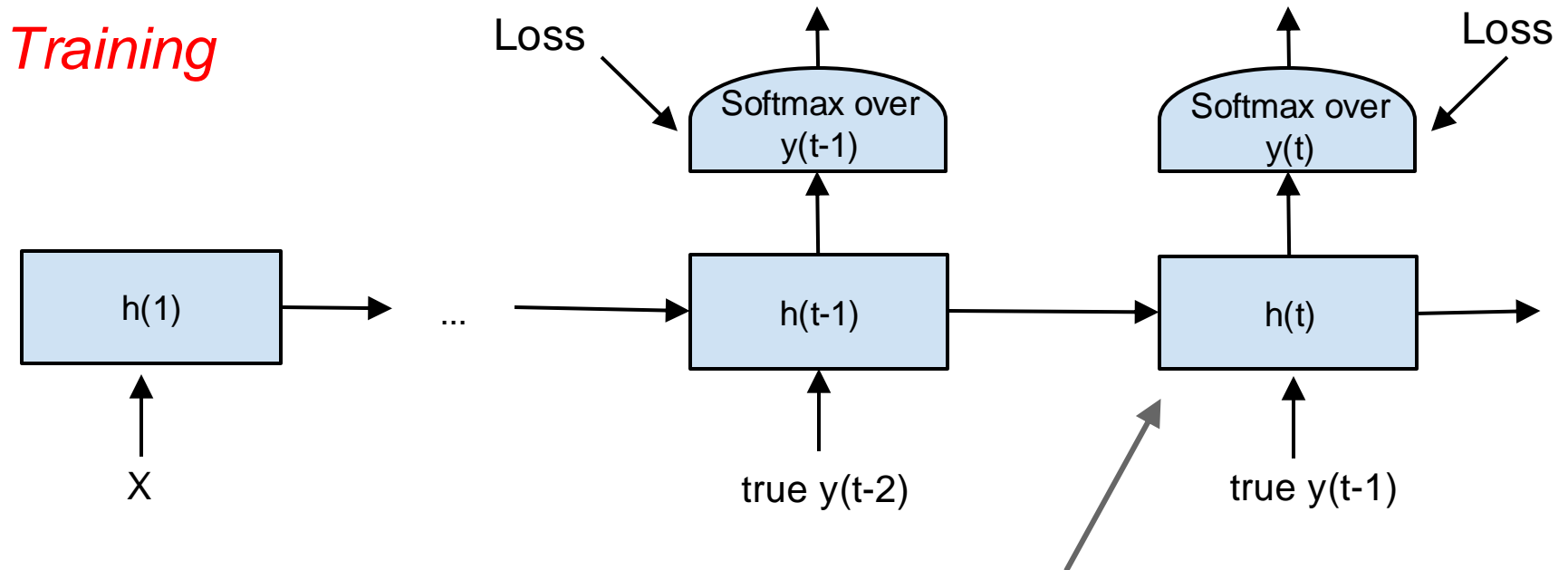
Cross Entropy (Negative Log Likelihood) Loss

- Log Likelihood, by chain rule is sum of next step log likelihoods

$$\log p(\mathbf{y}|\mathbf{x}) = \sum_{i=1}^N \log p(y_i | y_{<i}, \mathbf{x})$$

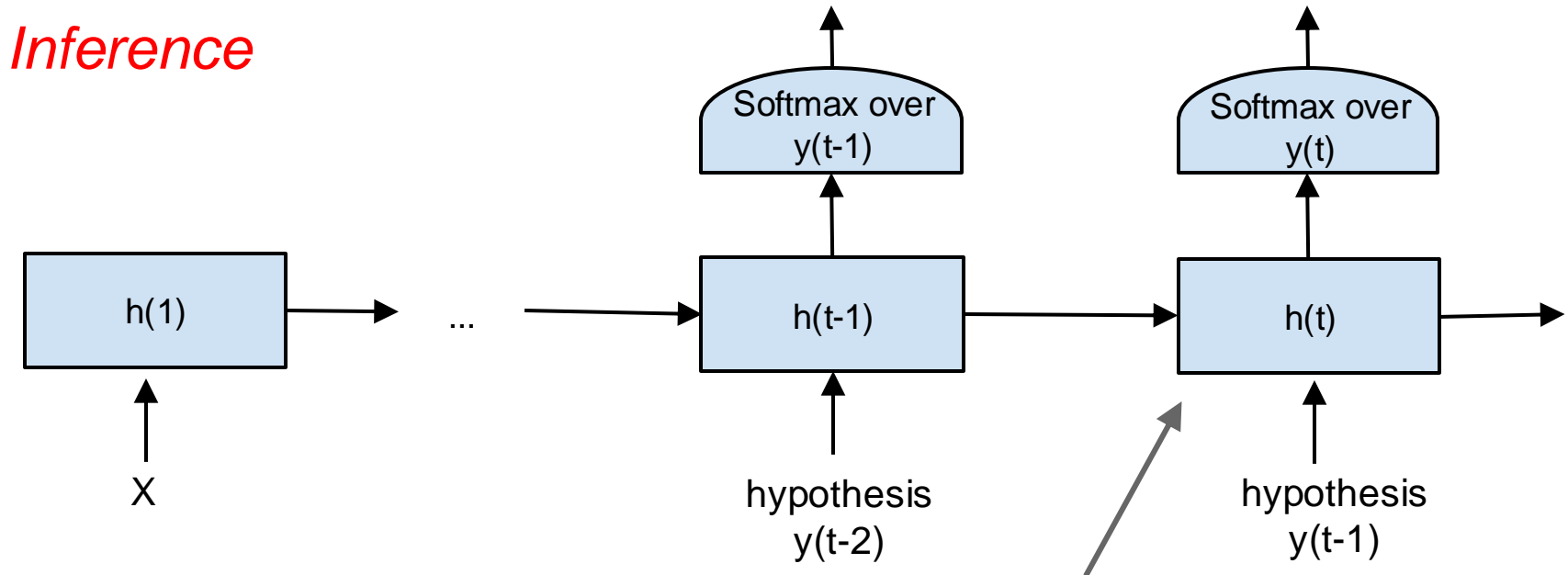
- Supervised classification for each time step
 - depends on input, past outputs, which are known during training

Training and Inference Mismatch



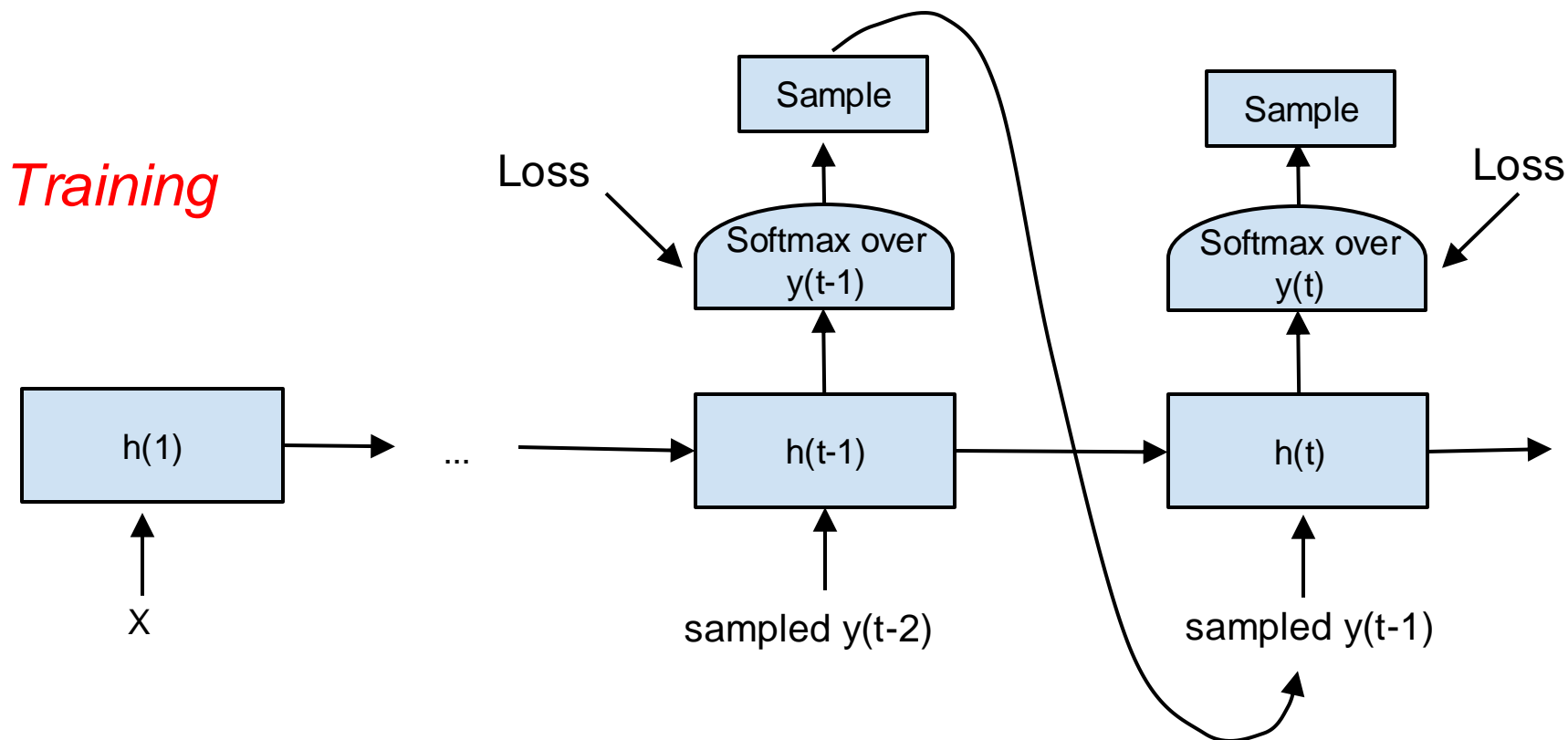
$$P(y_t|h_t) \text{ with } h_t = f(h_{t-1}, y_{t-1}; \theta)$$

Training and Inference Mismatch



$$P(y_t|h_t) \text{ with } h_t = f(h_{t-1}, y_{t-1}; \theta)$$

Scheduled Sampling



$$P(y_t|h_t) \text{ with } h_t = f(h_{t-1}, \hat{y}_{t-1}; \theta)$$

Scheduled Sampling

Machine Translation Model	Bleu-4	Meteor	Cider
Baseline	28.8	24.2	89.5
Baseline with dropout	28.1	23.9	87.0
Scheduled sampling	30.6	24.3	92.1

Parsing Model	F1
Baseline LSTM with dropout	87.00
Scheduled sampling with dropout	88.68

Speech Recognition Model	WER
LAS + LM Rescoring	12.6
LAS + Sampling + LM Rescoring	10.3

New Architectures