10707
Deep Learning

Russ Salakhutdinov

Machine Learning Department
rsalakhu@cs.cmu.edu

Sequence Model / Transformers

Slides borrowed from ICML Tutorial

Seq2Seq ICML Tutorial

Oriol Vinyals and Navdeep Jaitly

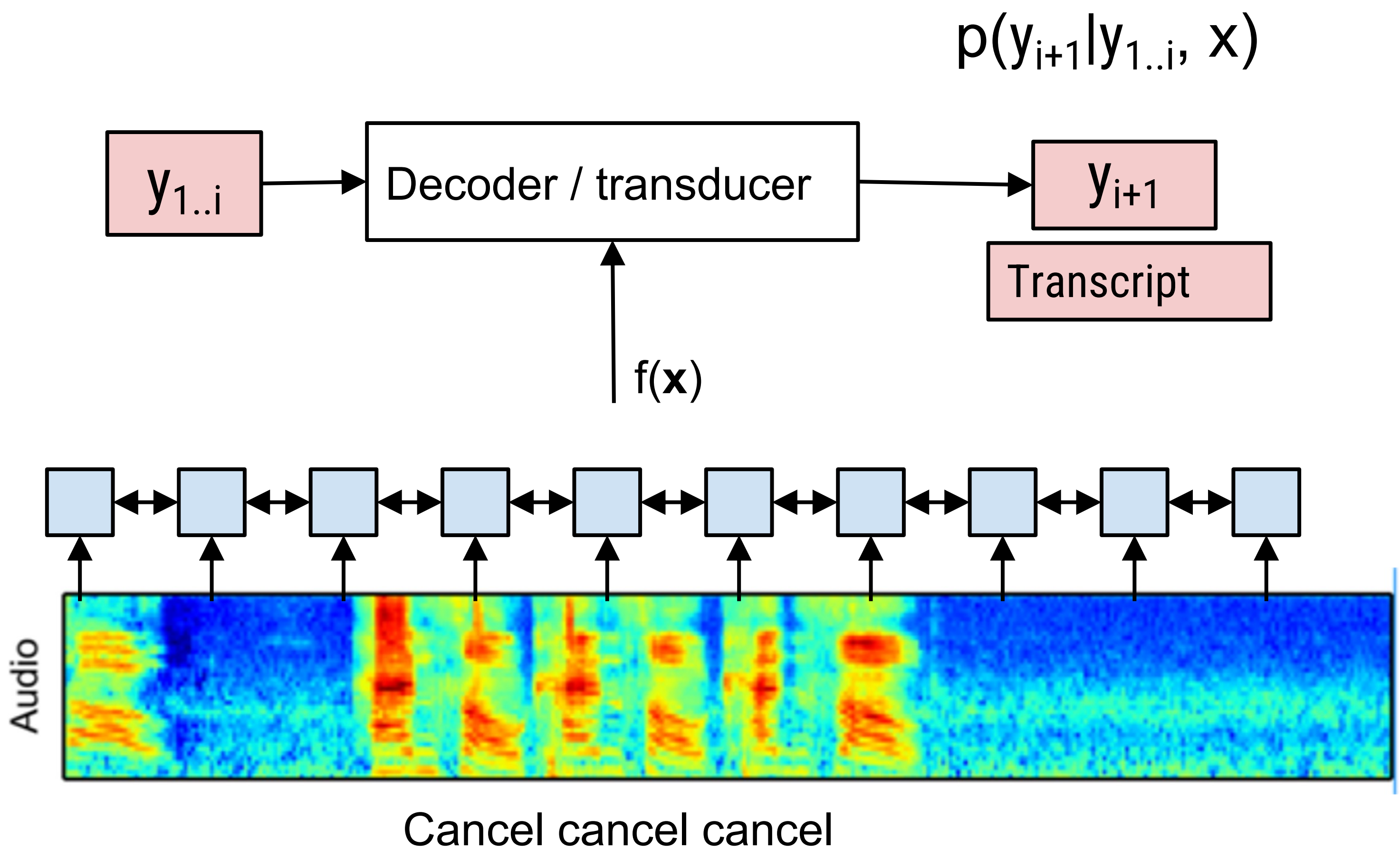
@OriolVinyalsML | @NavdeepLearning

Site: https://sites.google.com/view/seq2seq-icml17

Sydney, Australia, 2017

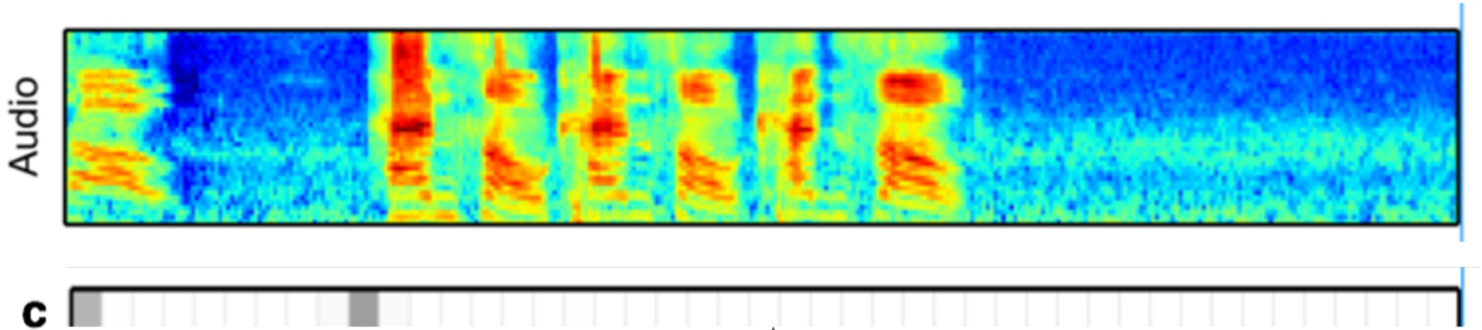
Applications

Speech Recognition



Chan, W., Jaitly, N., Le, Q., Vinyals, O. "*Listen Attend and Spell.*" *ICASSP (2015).*

Attention Example



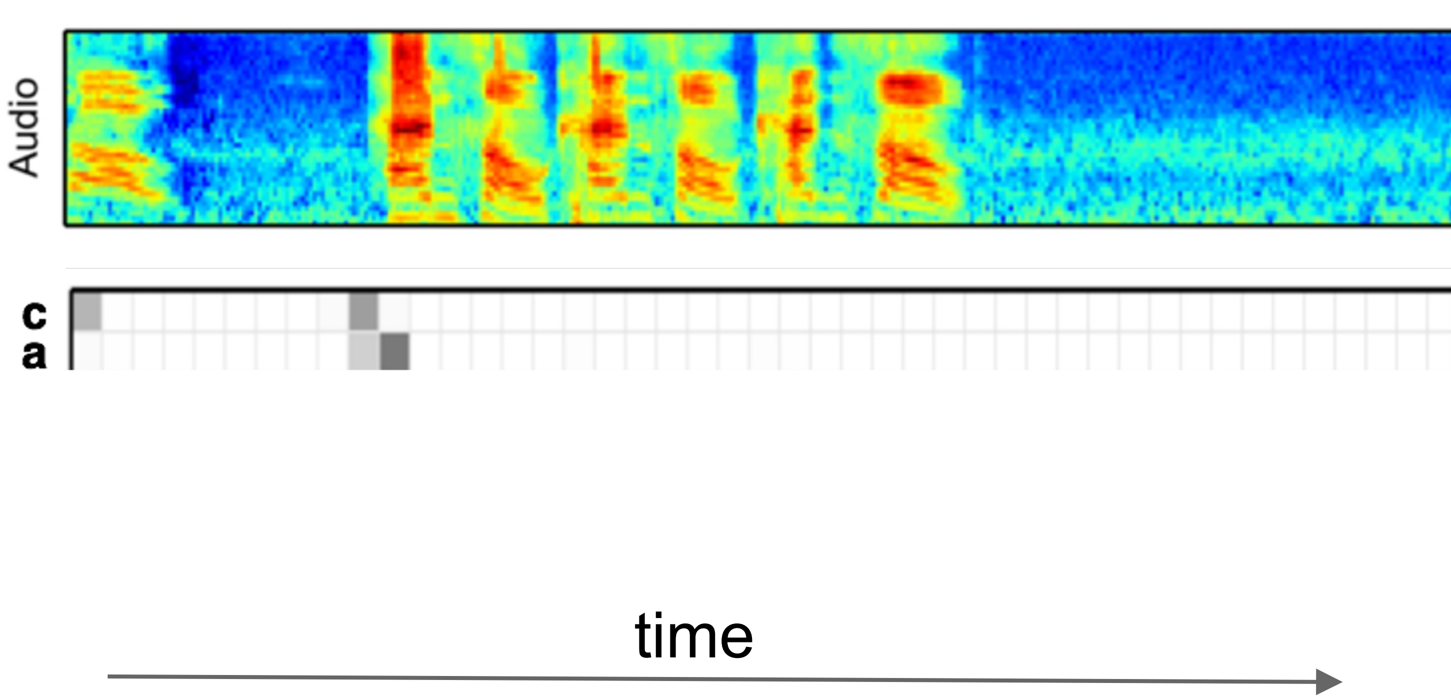
prediction derived from “attending” to segment of input

Attention vector - where the model thinks the relevant information is to be found

time

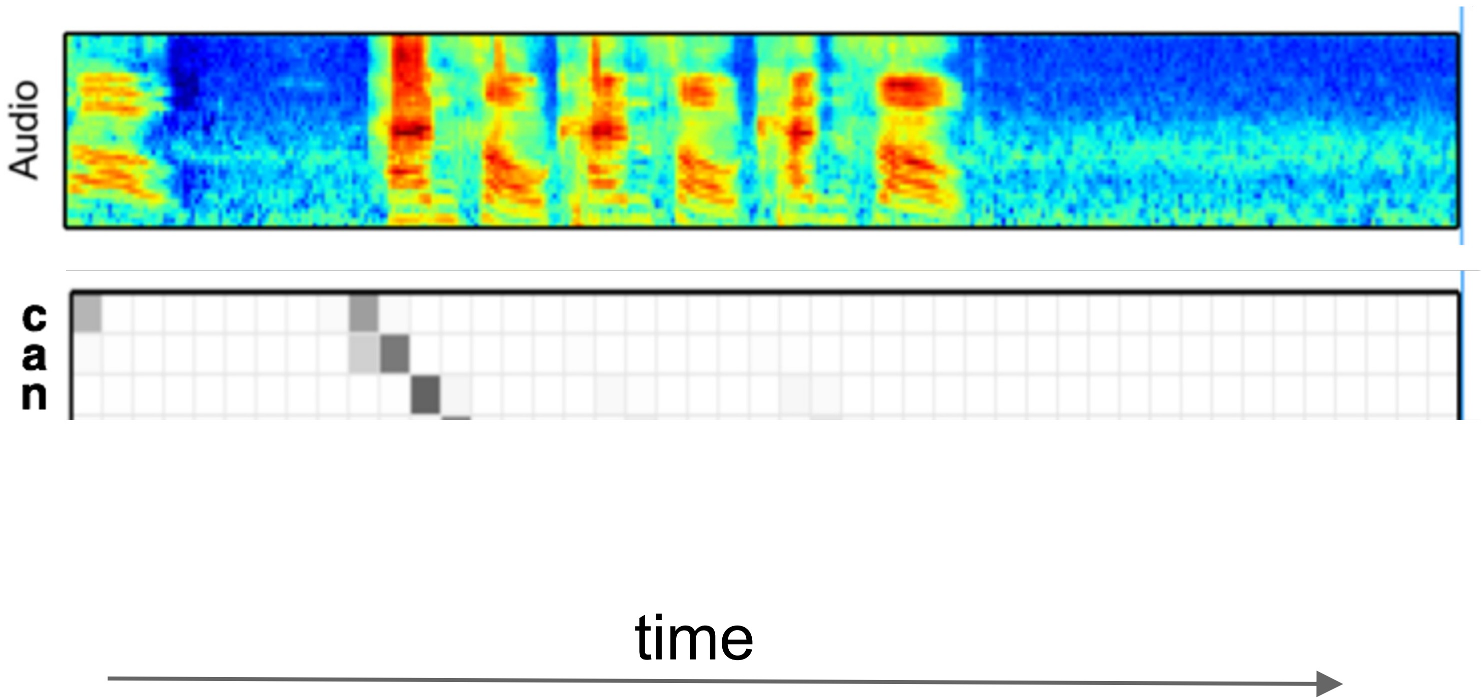
Chan, W., Jaitly, N., Le, Q., Vinyals, O. “*Listen Attend and Spell*.” *ICASSP (2015).*

Attention Example



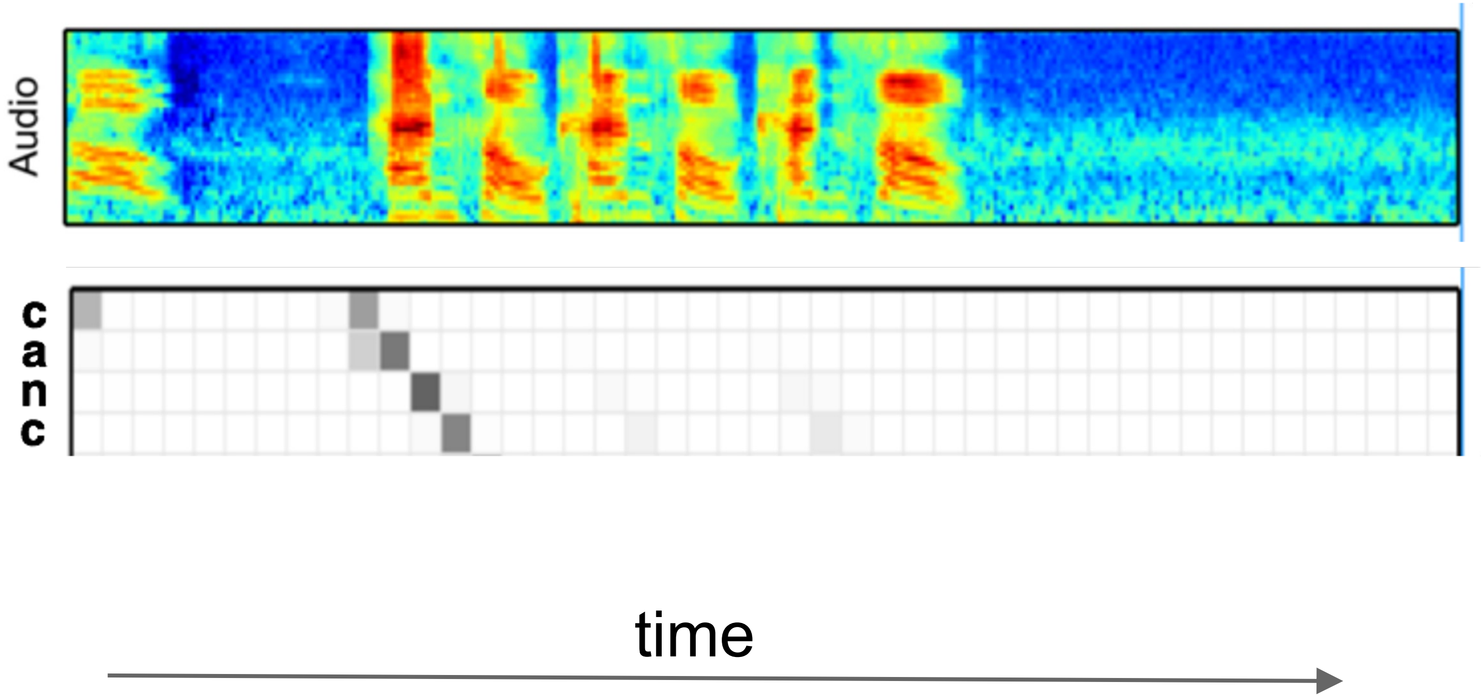
Chan, W., Jaitly, N., Le, Q., Vinyals, O. "*Listen Attend and Spell.*" *ICASSP (2015).*

Attention Example



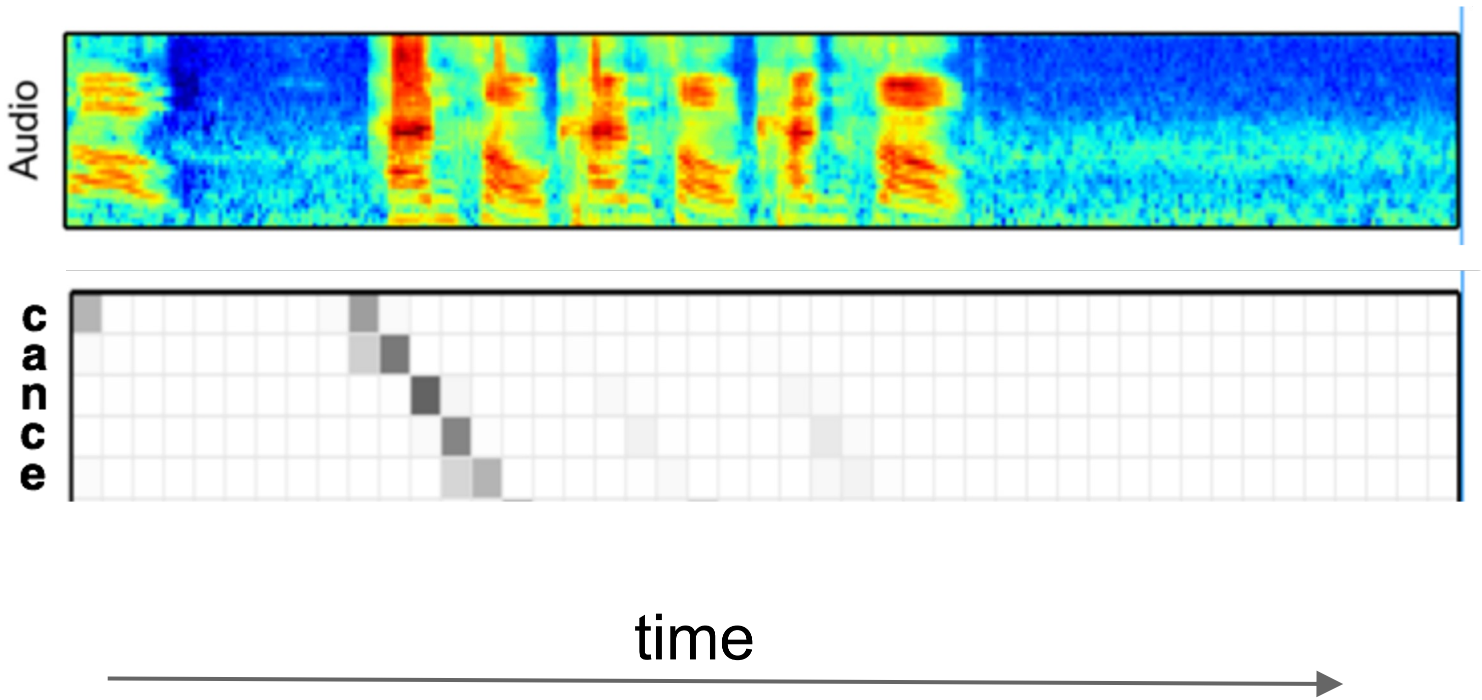
Chan, W., Jaitly, N., Le, Q., Vinyals, O. “*Listen Attend and Spell.*” *ICASSP (2015).*

Attention Example



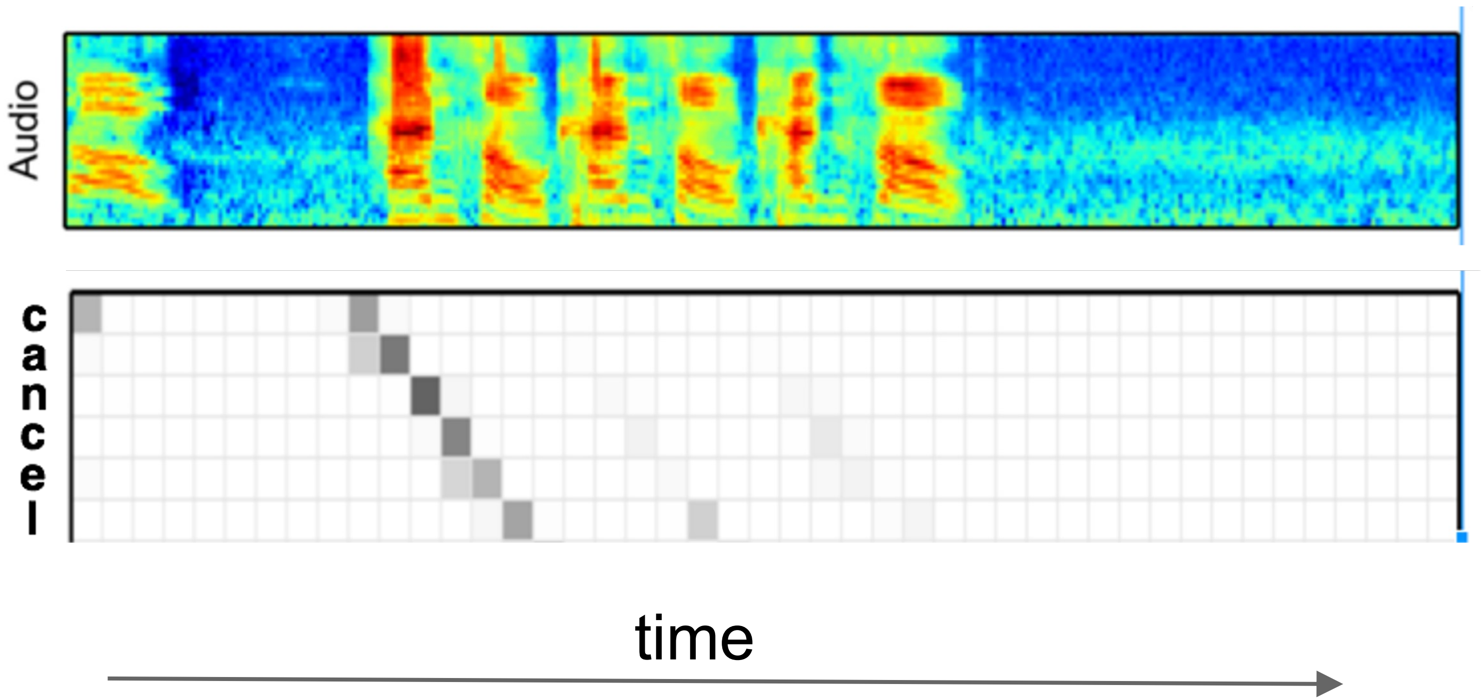
Chan, W., Jaitly, N., Le, Q., Vinyals, O. "*Listen Attend and Spell.*" *ICASSP (2015).*

Attention Example



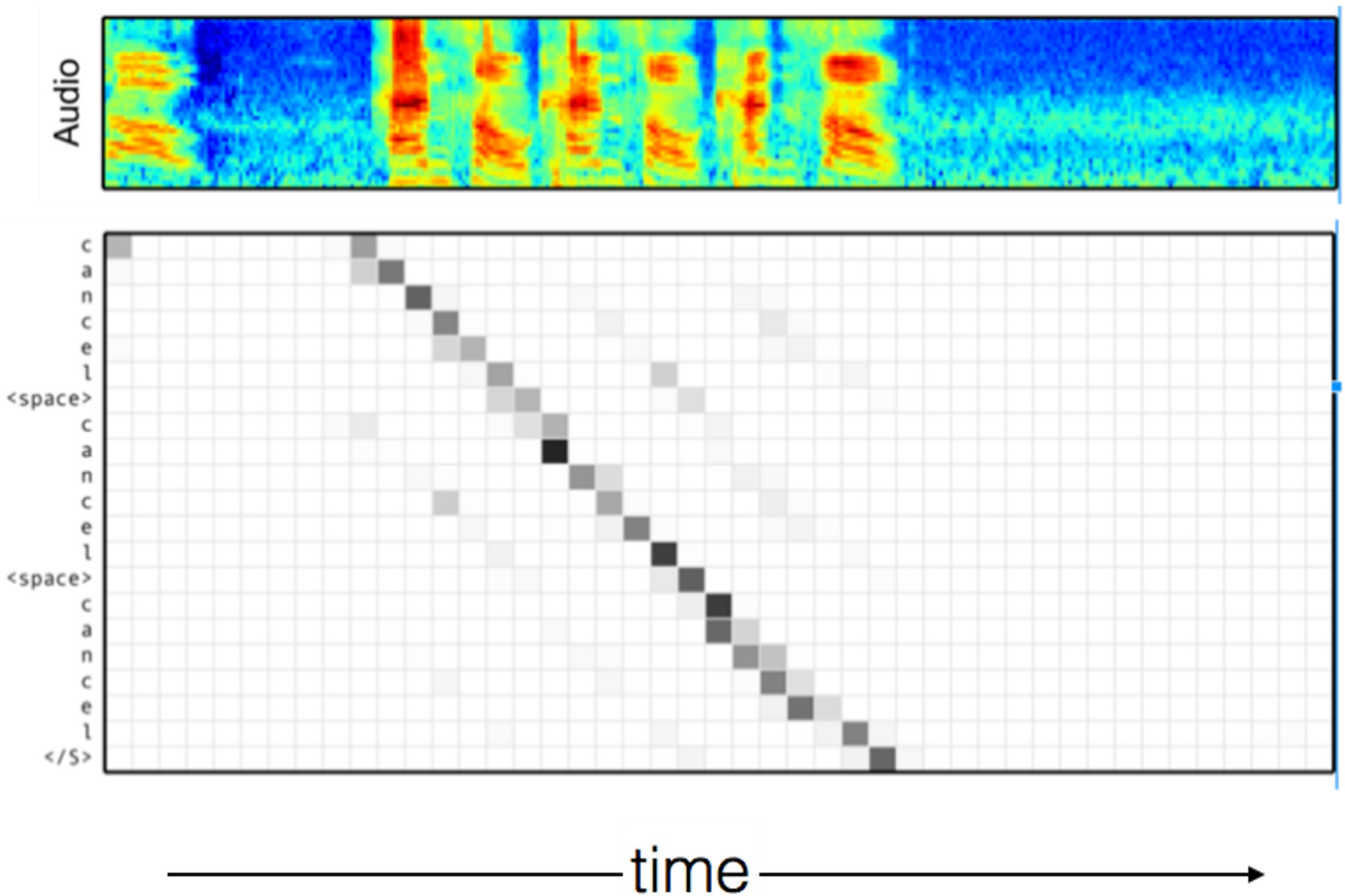
Chan, W., Jaitly, N., Le, Q., Vinyals, O. “*Listen Attend and Spell*.” *ICASSP (2015).*

Attention Example



Chan, W., Jaitly, N., Le, Q., Vinyals, O. "*Listen Attend and Spell.*" *ICASSP (2015).*

Attention Example



Chan, W., Jaitly, N., Le, Q., Vinyals, O. "*Listen Attend and Spell*." *ICASSP (2015).*

Caption Generation with Visual Attention

A man riding a horse in a field.

Xu et al, ICML 2015

Caption Generation with Visual Attention

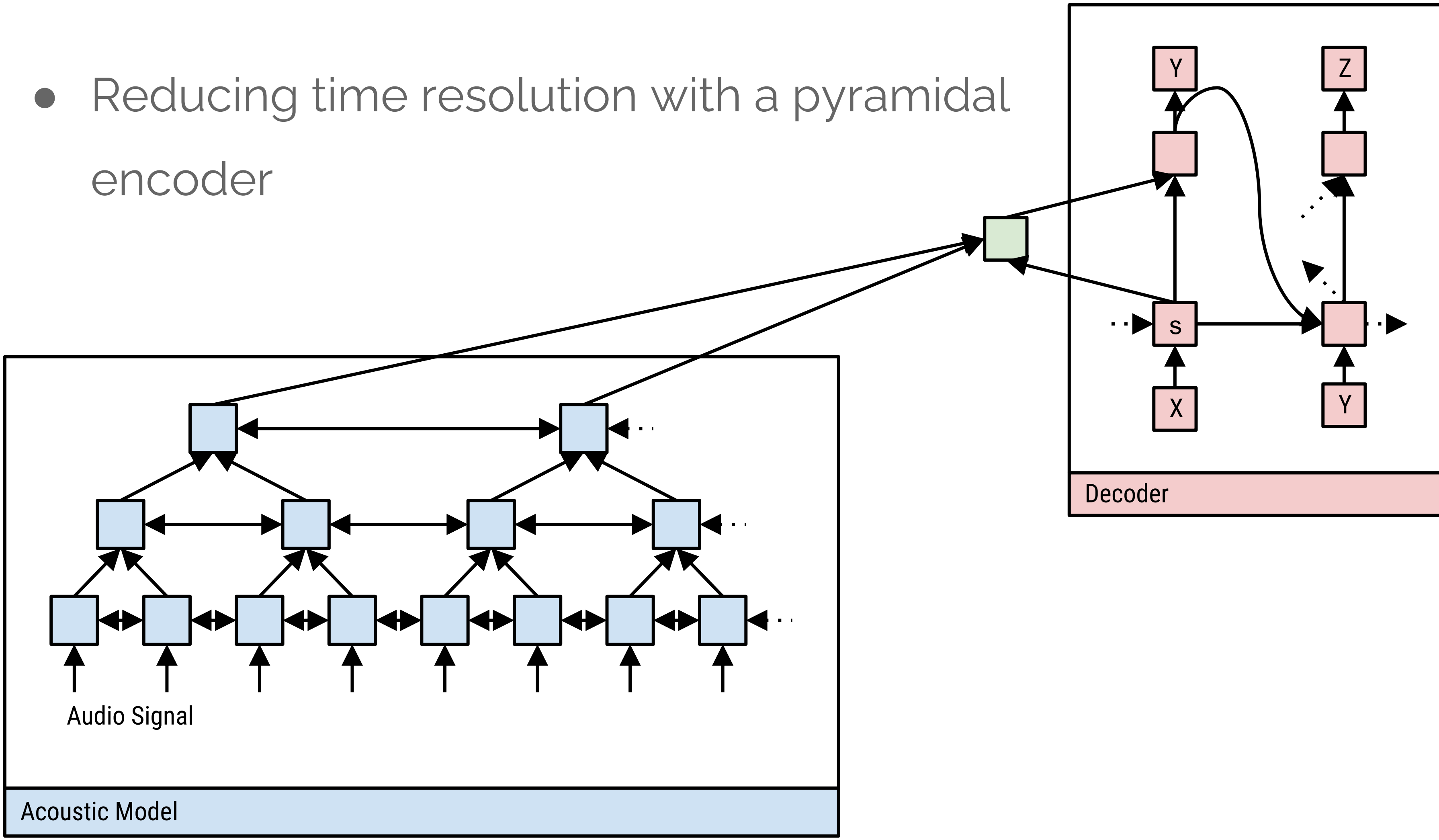
A woman holding a clock in her hand.

A large white bird standing in a forest.

Xu et al, ICML 2015

Listen Attend and Spell (LAS)

- Reducing time resolution with a pyramidal encoder



Chan, W., Jaitly, N., Le, Q., Vinyals, O. “*Listen Attend and Spell*.” *ICASSP (2015).*

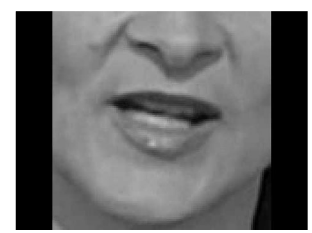
LAS Results

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

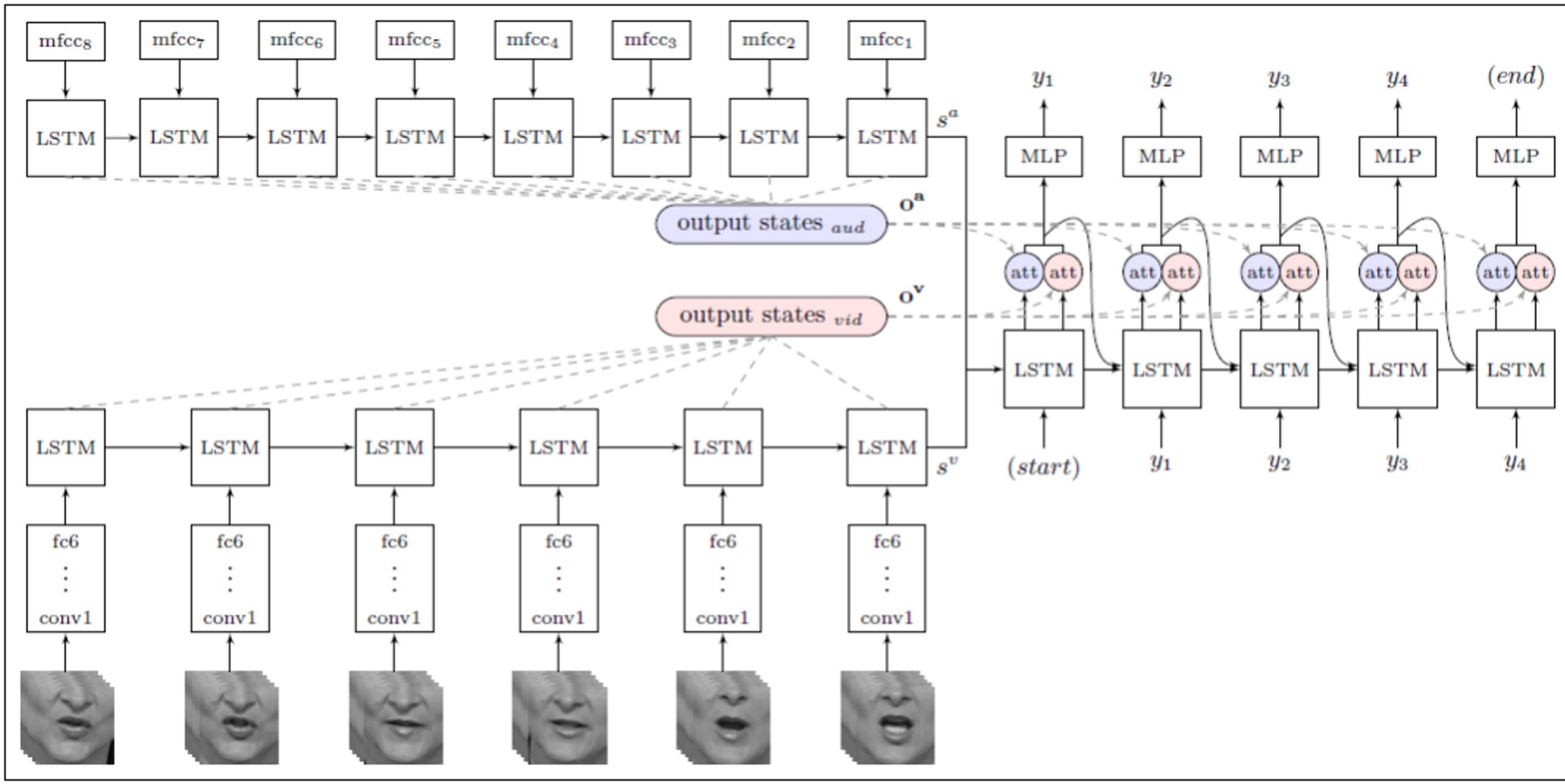
http://www.robots.ox.ac.uk/~vgg/data/lip_reading/



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

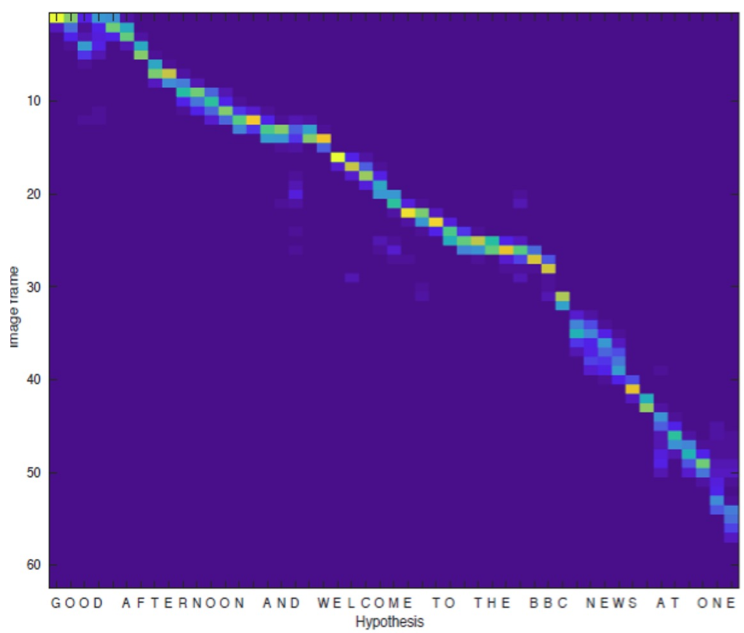
Separate embedding and attention for audio and visual streams

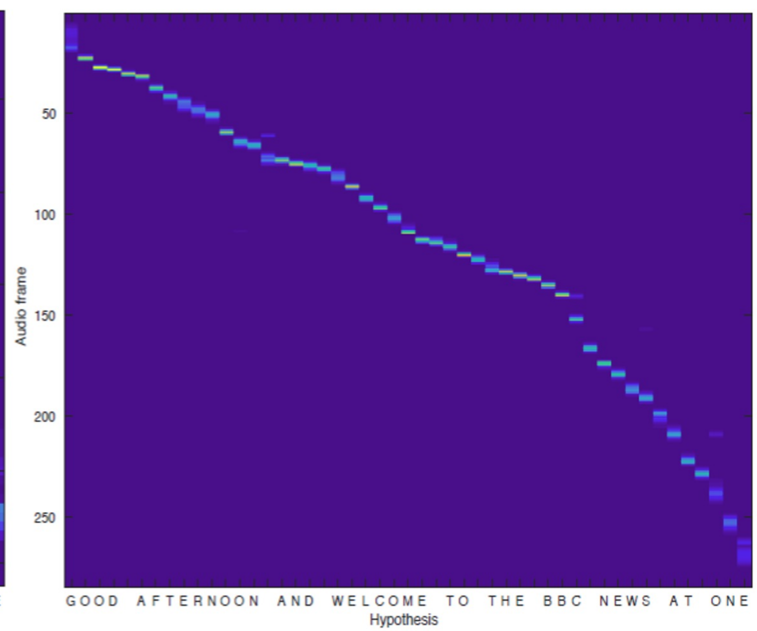


Chung, J., et al. "Lip reading sentences in the wild." *CVPR* (2017).

Lip Reading

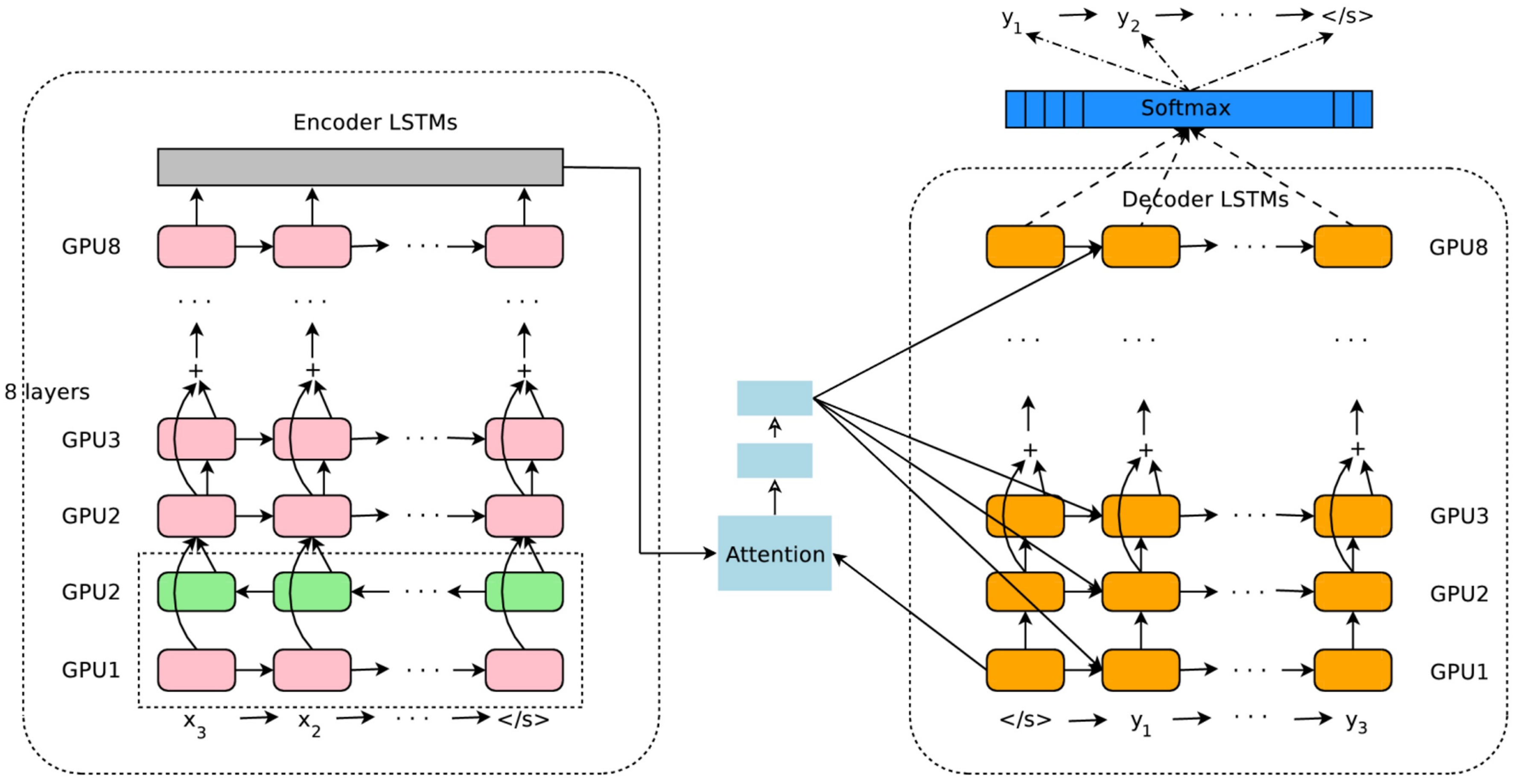
Method	SNR	CER	WER	BLEU†
Lips only				
Professional‡	-	58.7%	73.8%	23.8
WAS	-	59.9%	76.5%	35.6
WAS+CL	-	47.1%	61.1%	46.9
WAS+CL+SS	-	42.4%	58.1%	50.0
WAS+CL+SS+BS	-	39.5%	50.2%	54.9





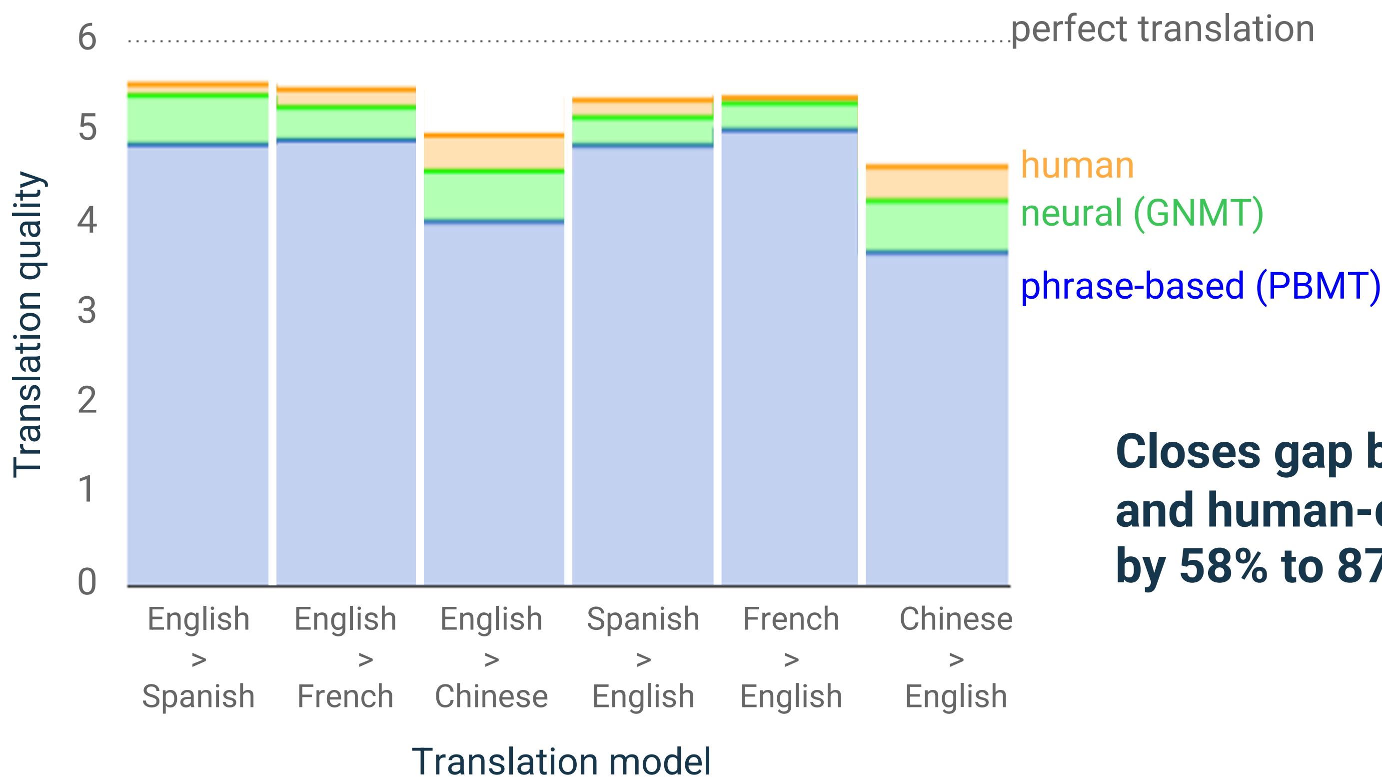
Chung, J., et al. "Lip reading sentences in the wild." *CVPR* (2017).

Google Neural Machine Translation System



Wu, Y., et al. "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." *arxiv (2016).*

Google Neural Machine Translation System



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

research.googleblog.com/2016/09/a-neural-network-for-machine.html

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).

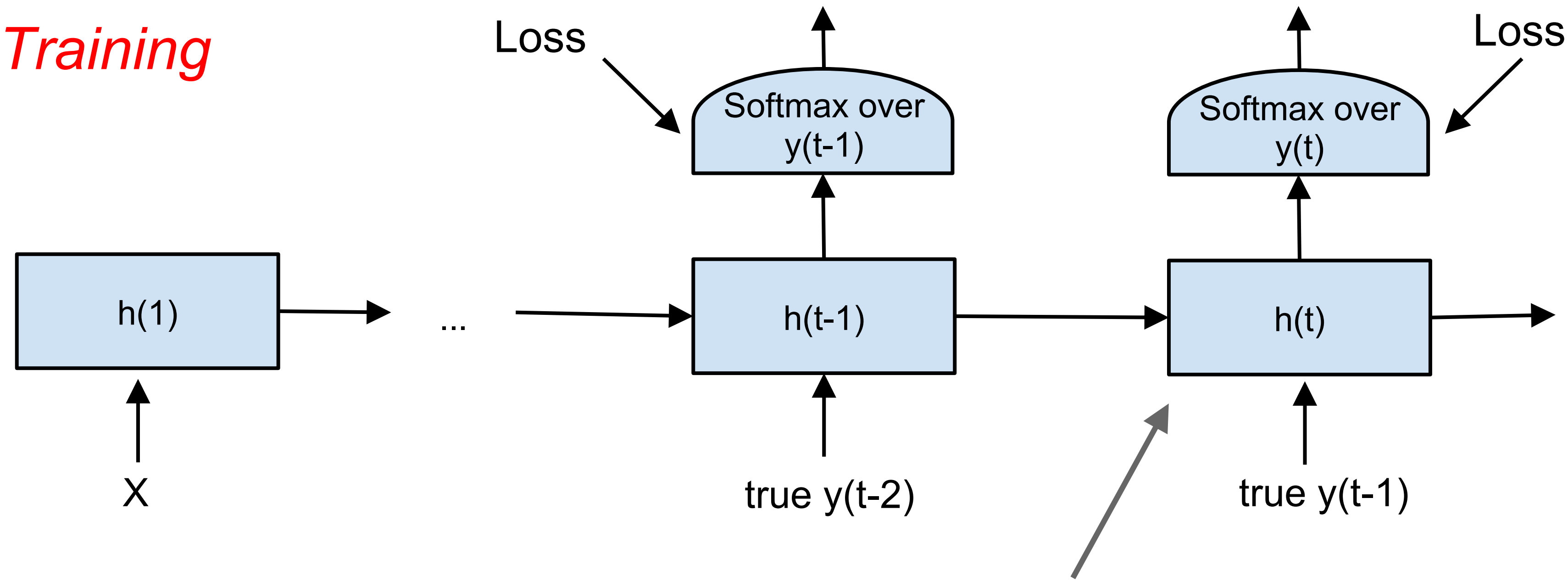
Cross Entropy (Negative Log Likelihood) Loss

- 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

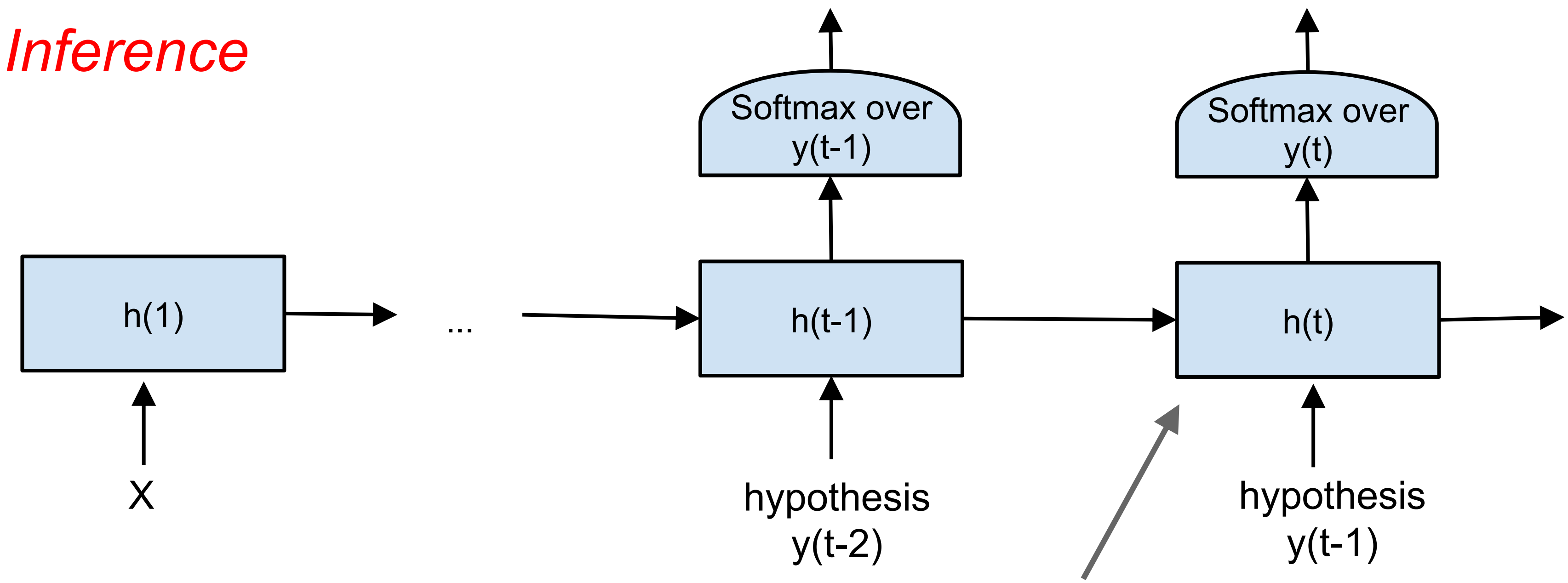
Training and Inference Mismatch



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

Bengio, S., et al. "Scheduled sampling for sequence prediction with recurrent neural networks." *NIPS (2015).*

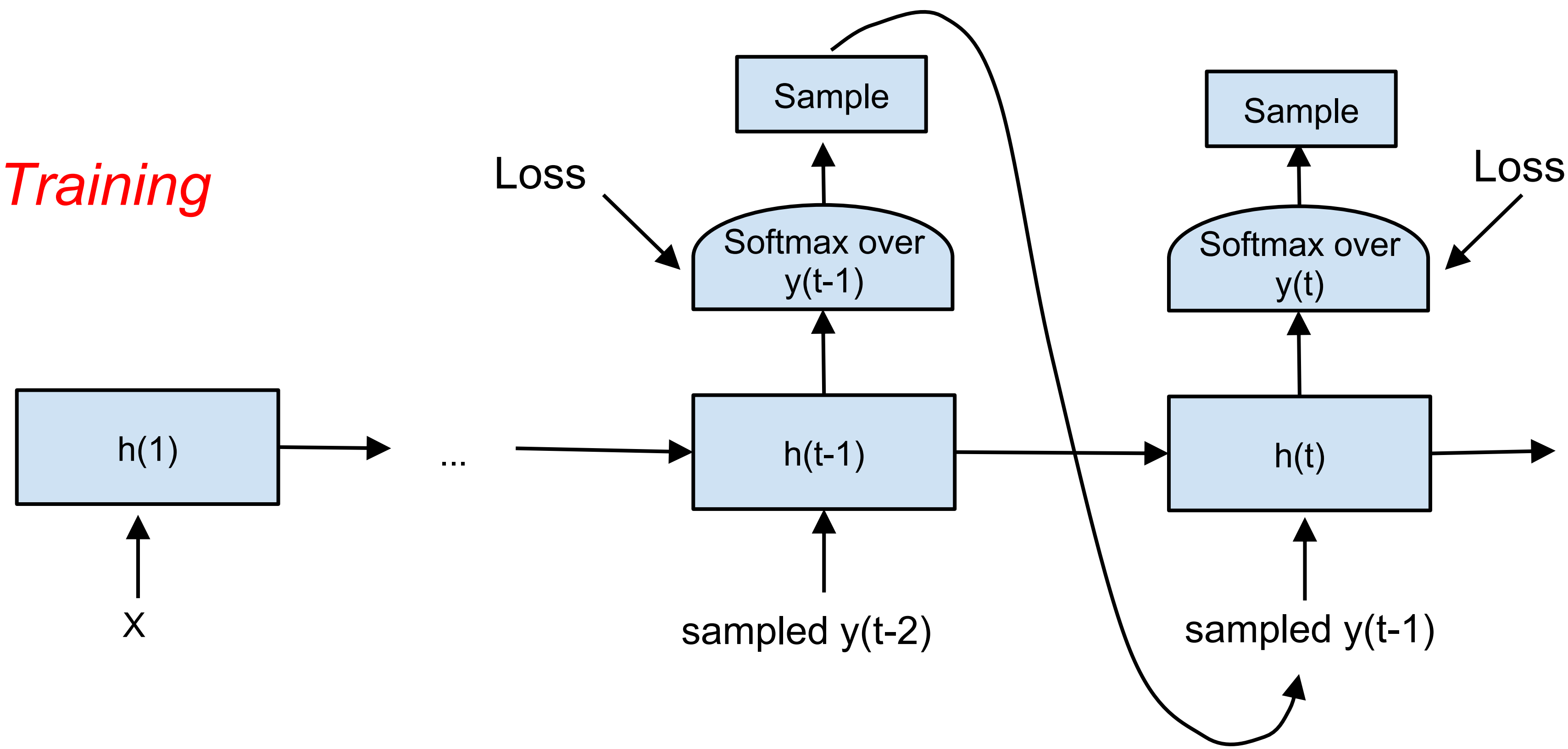
Training and Inference Mismatch



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

Bengio, S., et al. "Scheduled sampling for sequence prediction with recurrent neural networks." *NIPS (2015).*

Scheduled Sampling



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

Bengio, S., et al. “Scheduled sampling for sequence prediction with recurrent neural networks.” *NIPS (2015).*

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**

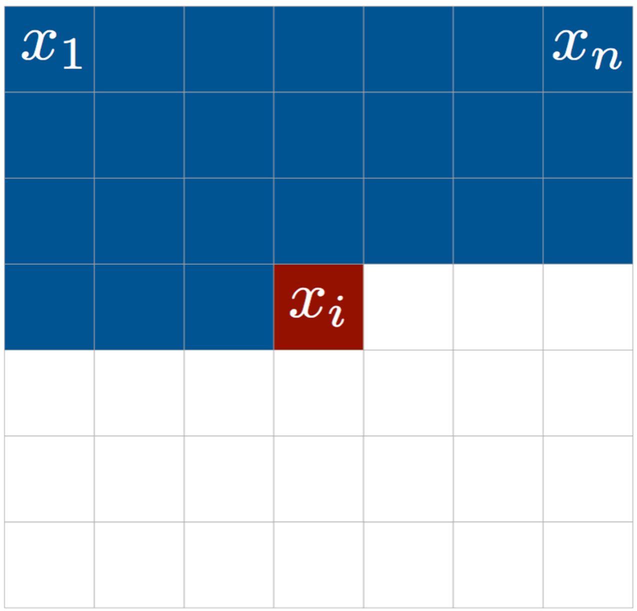
Parsing Model	F1
Baseline LSTM with dropout	87.00
Scheduled sampling with dropout	**88.68**

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

Bengio, S., et al. "Scheduled sampling for sequence prediction with recurrent neural networks." *NIPS (2015).*

Autoregressive Generative Models

Pixel RNN Model

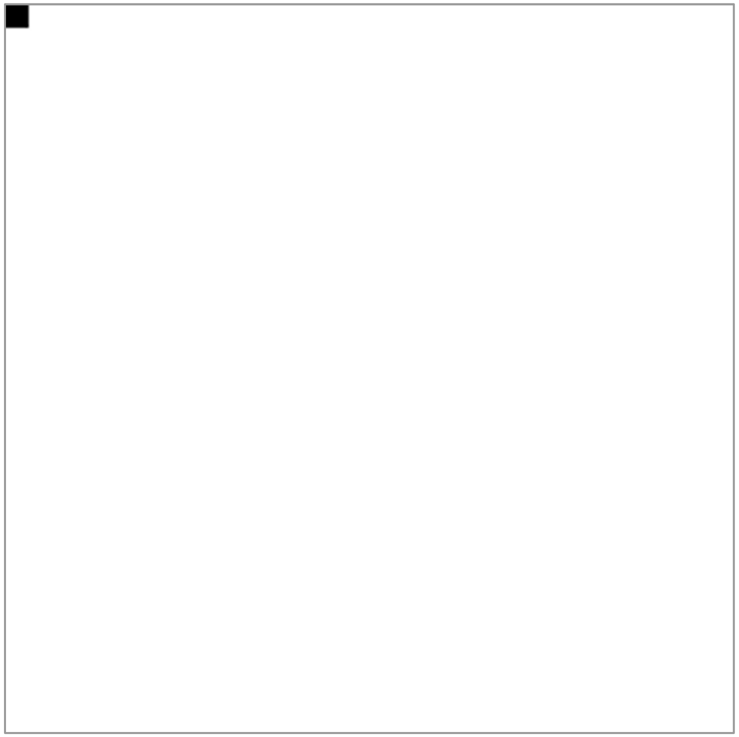


$$p(\mathbf{x}) = \prod_{i=1}^{n^2} p(x_i | x_1, ..., x_{i-1})$$

- Fully visible
- Similar to language models with RNNs
- Model pixels with Softmax

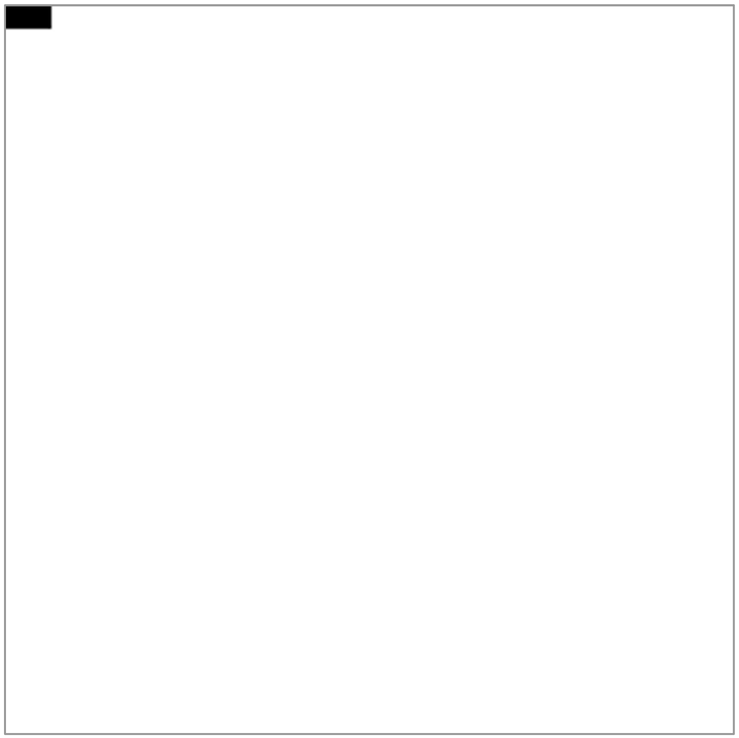
van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

Softmax Sampling



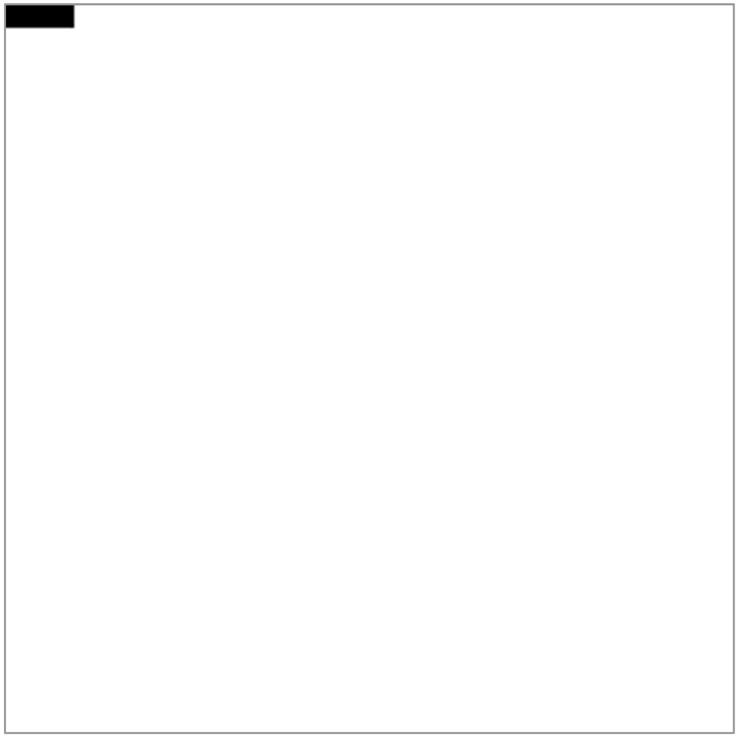
van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

Softmax Sampling



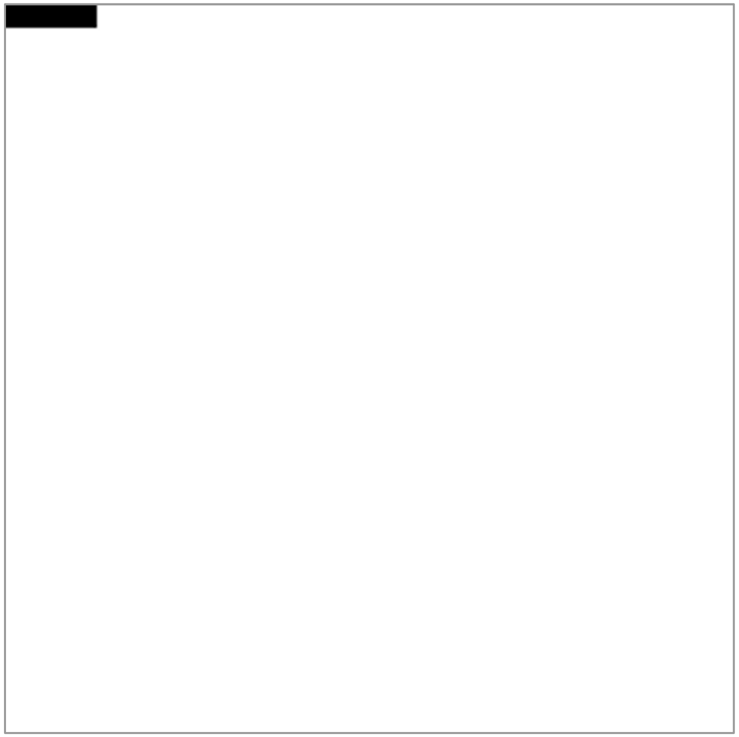
van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

Softmax Sampling



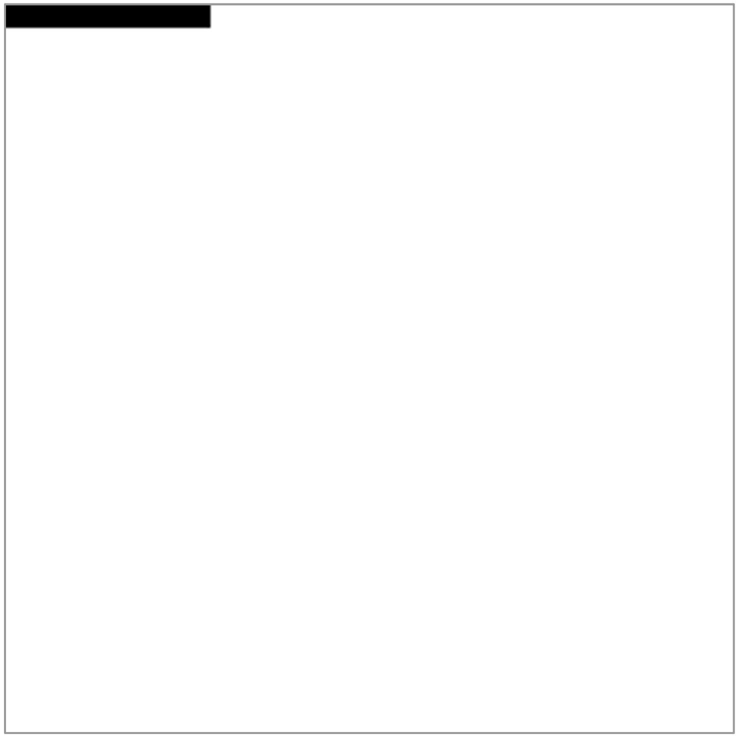
van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

Softmax Sampling



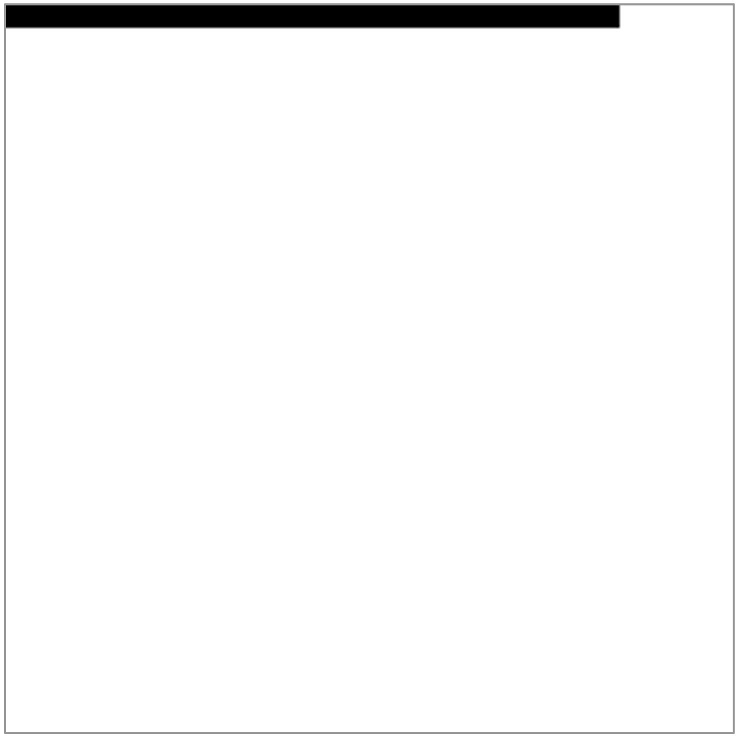
van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

Softmax Sampling



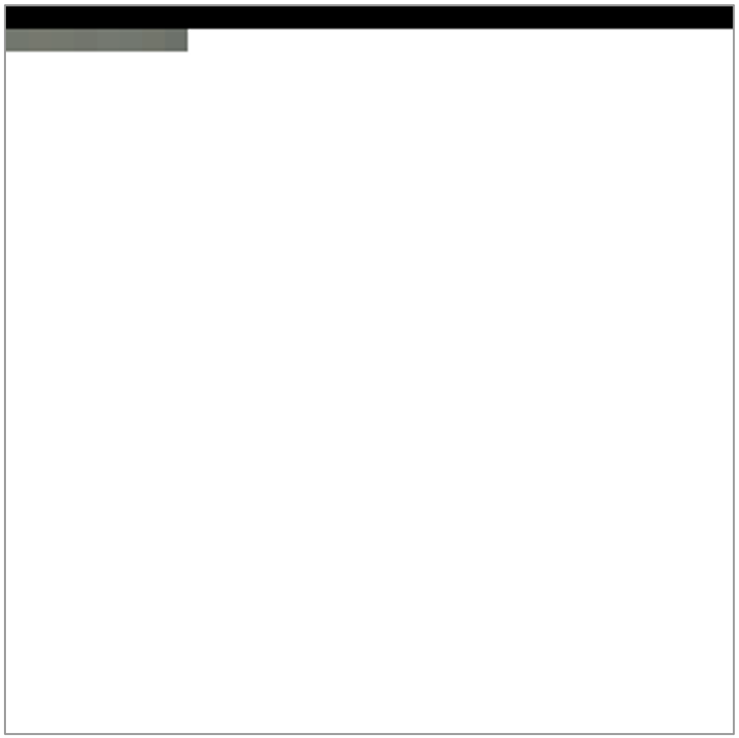
van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

Softmax Sampling



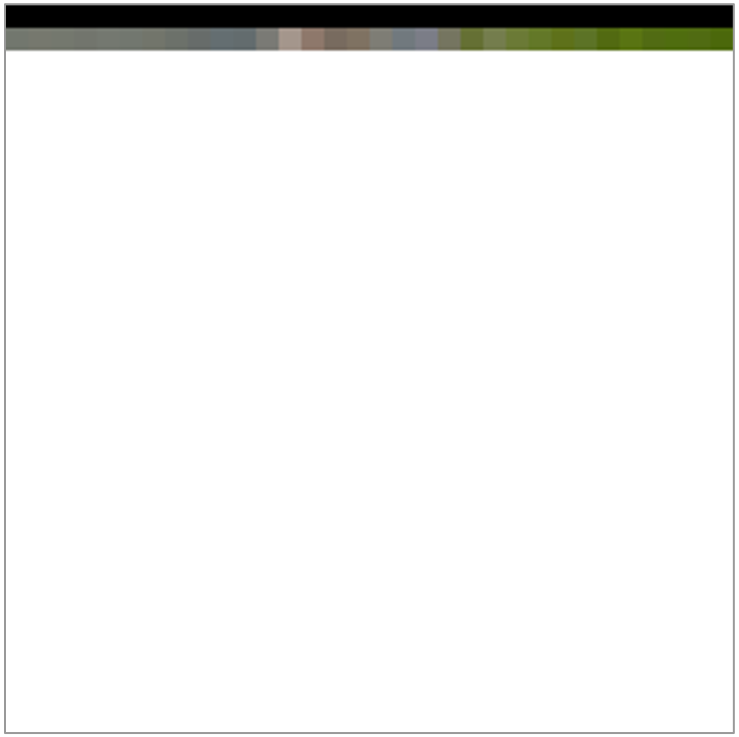
van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

Softmax Sampling



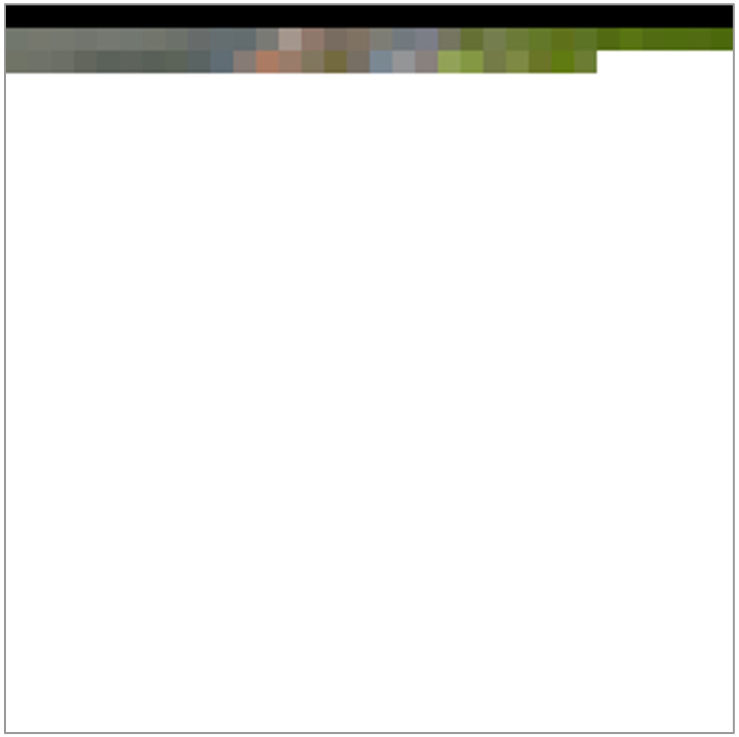
van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

Softmax Sampling



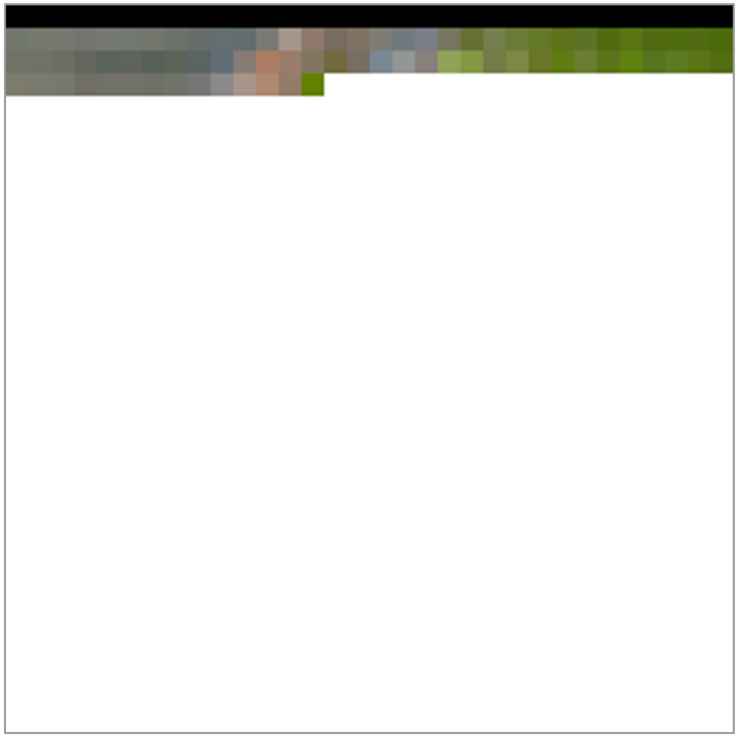
van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

Softmax Sampling



van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

Softmax Sampling

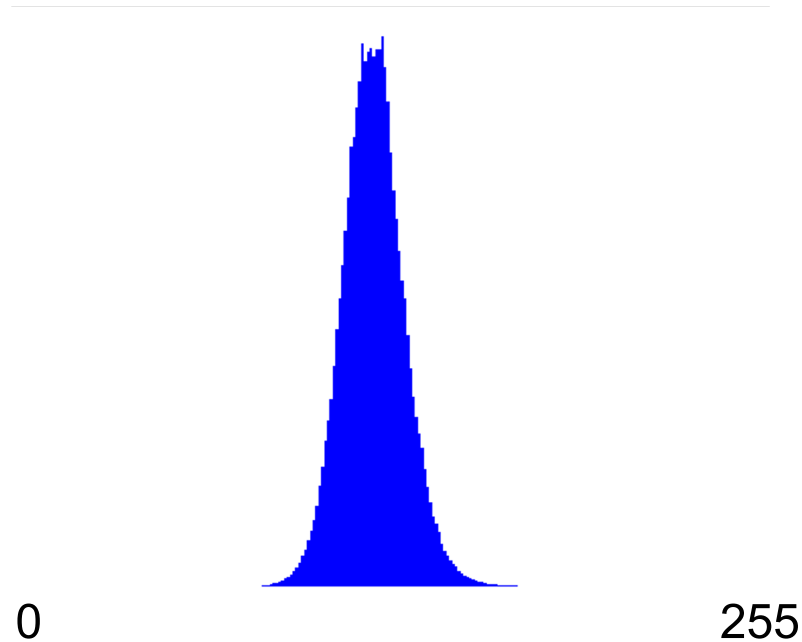


van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

Softmax Sampling

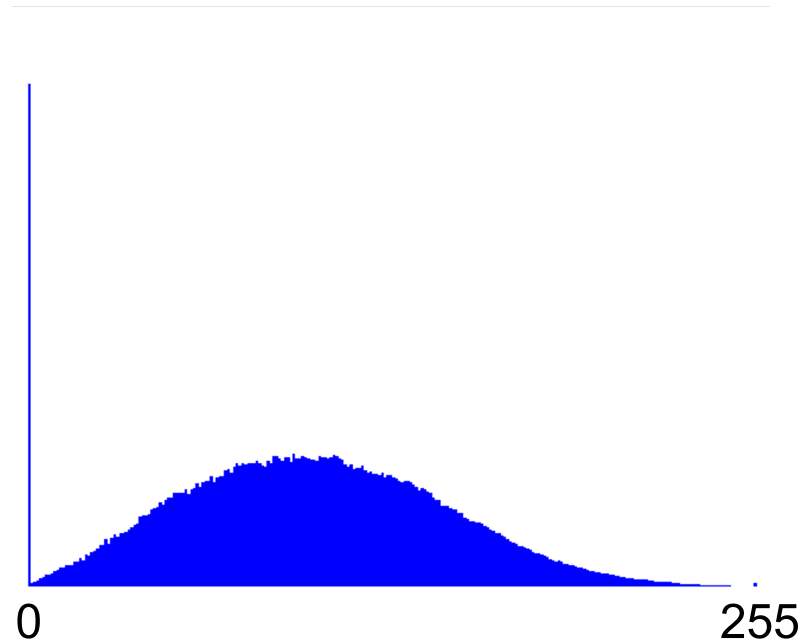
van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

Softmax Sampling



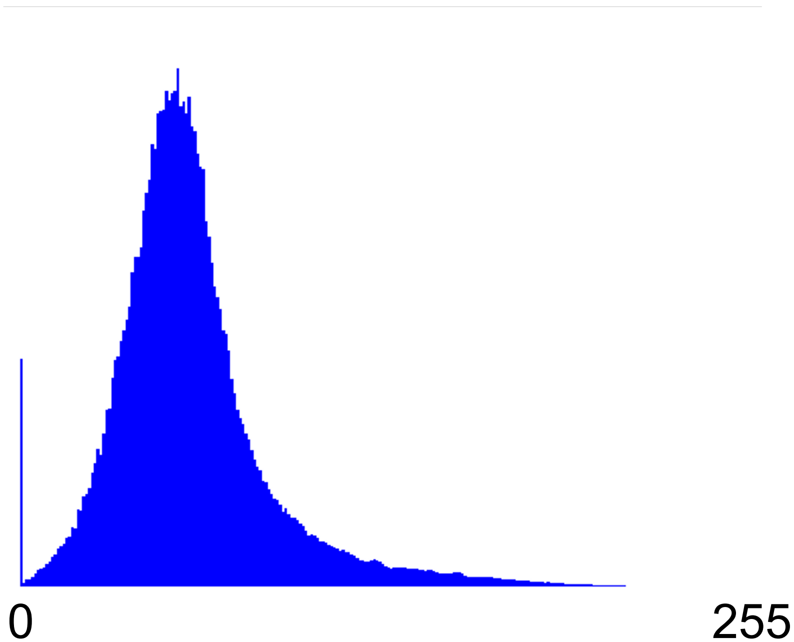
van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

Softmax Sampling



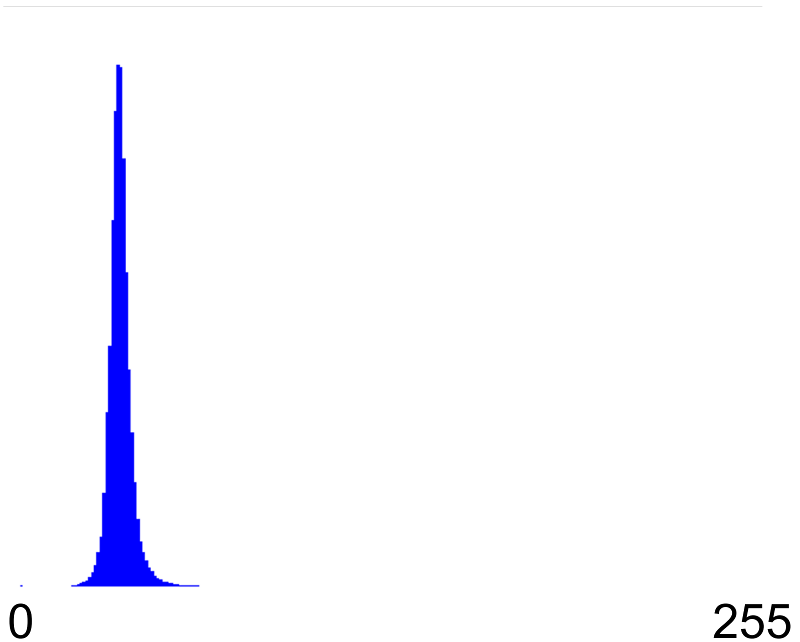
van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

Softmax Sampling



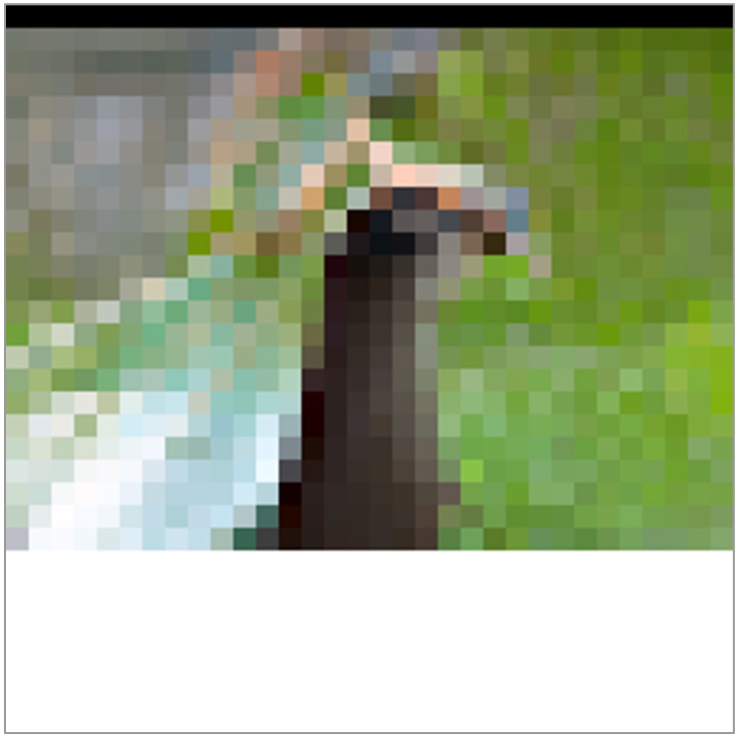
van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

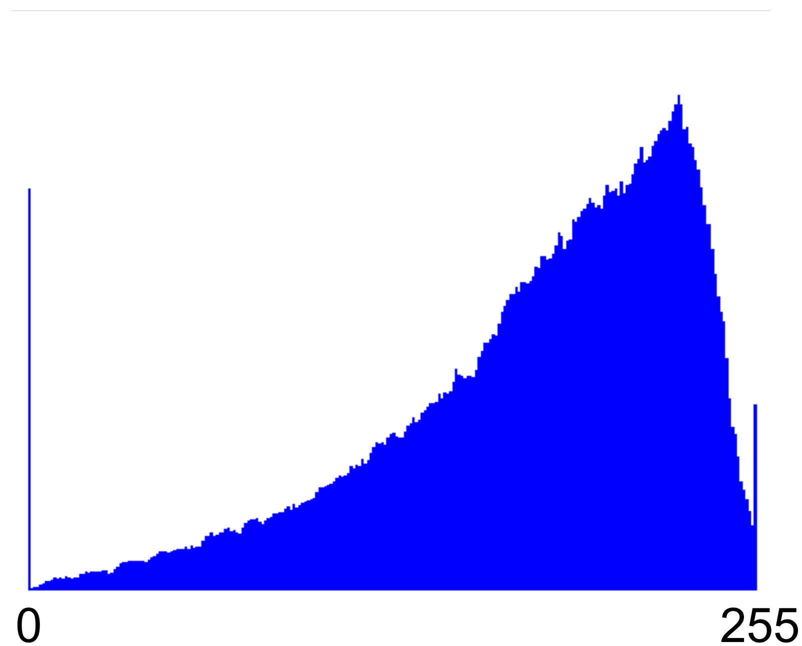
Softmax Sampling



van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

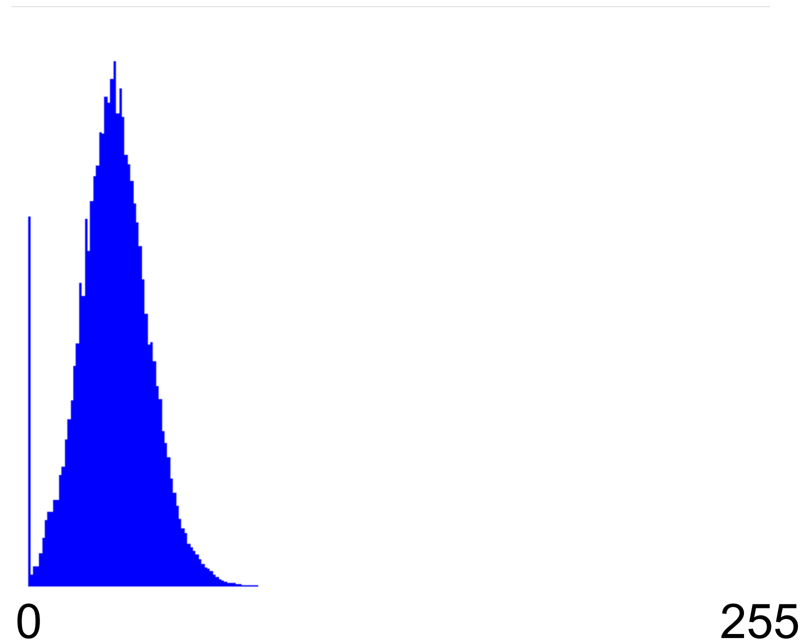
Softmax Sampling





van den Oord, A., et al. "Pixel Recurrent Neural Networks." *ICML (2016).*

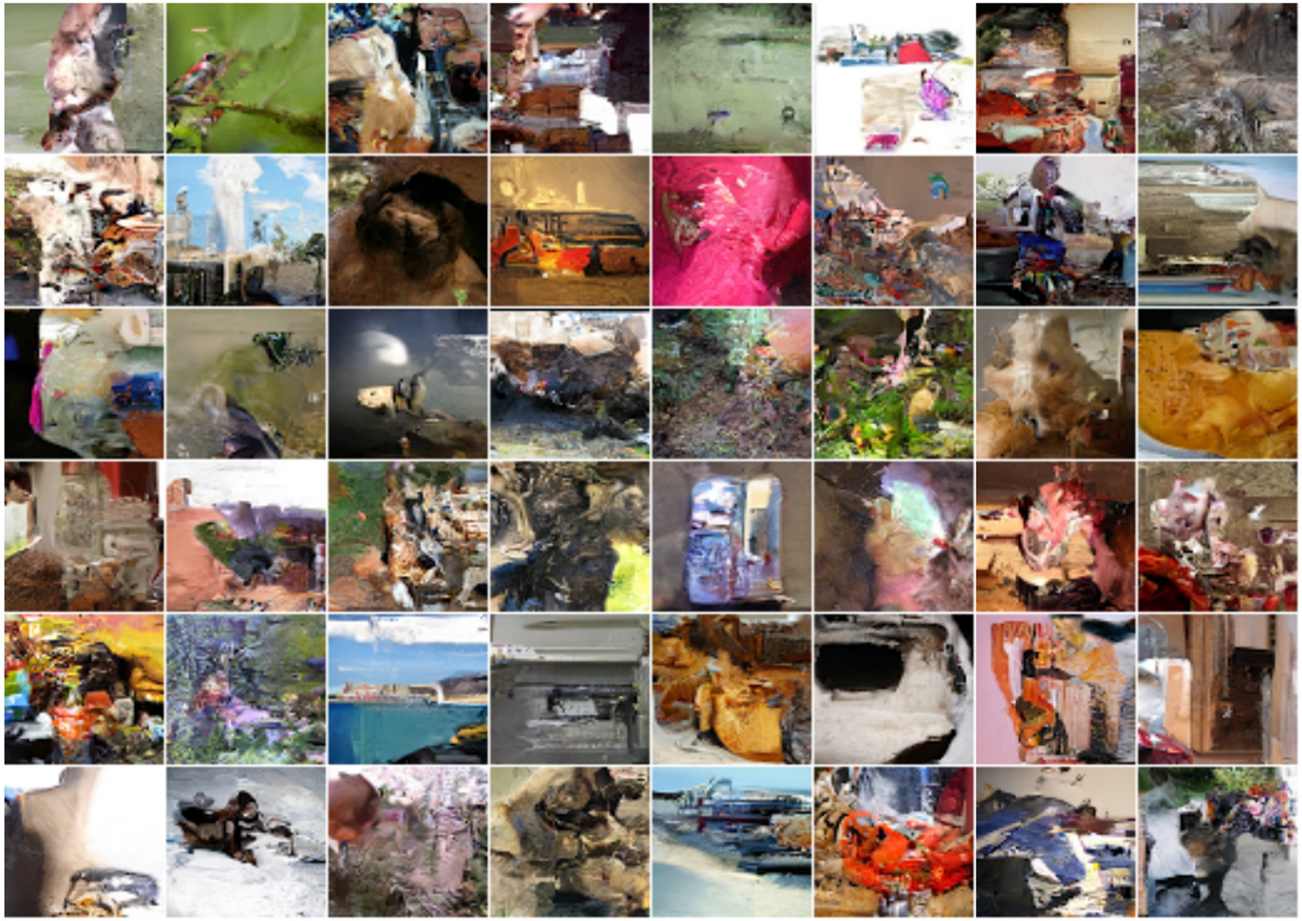
Softmax Sampling



van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

Pixel RNN

Sequence of Words == Sequence of Pixels



van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

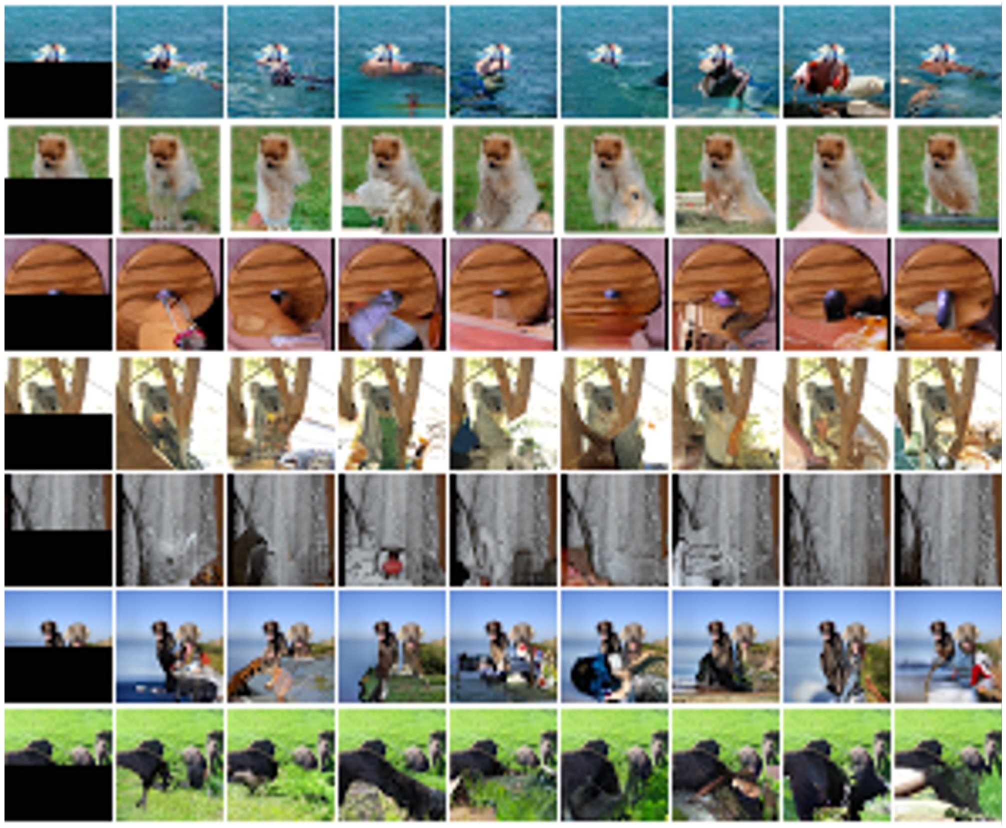
Pixel RNN

occluded

van den Oord, A., et al. "Pixel Recurrent Neural Networks." *ICML (2016).*

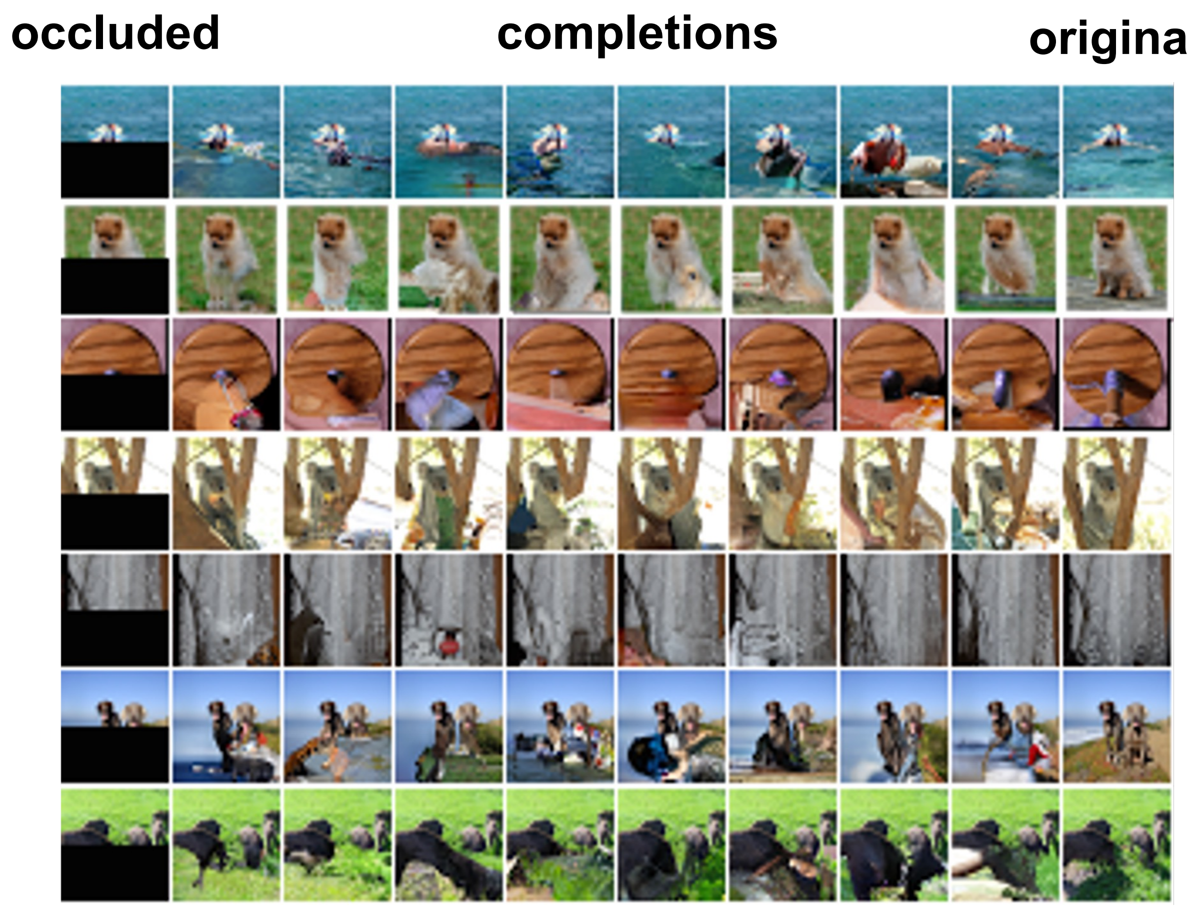
Pixel RNN

occluded **completions**



van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

Pixel RNN

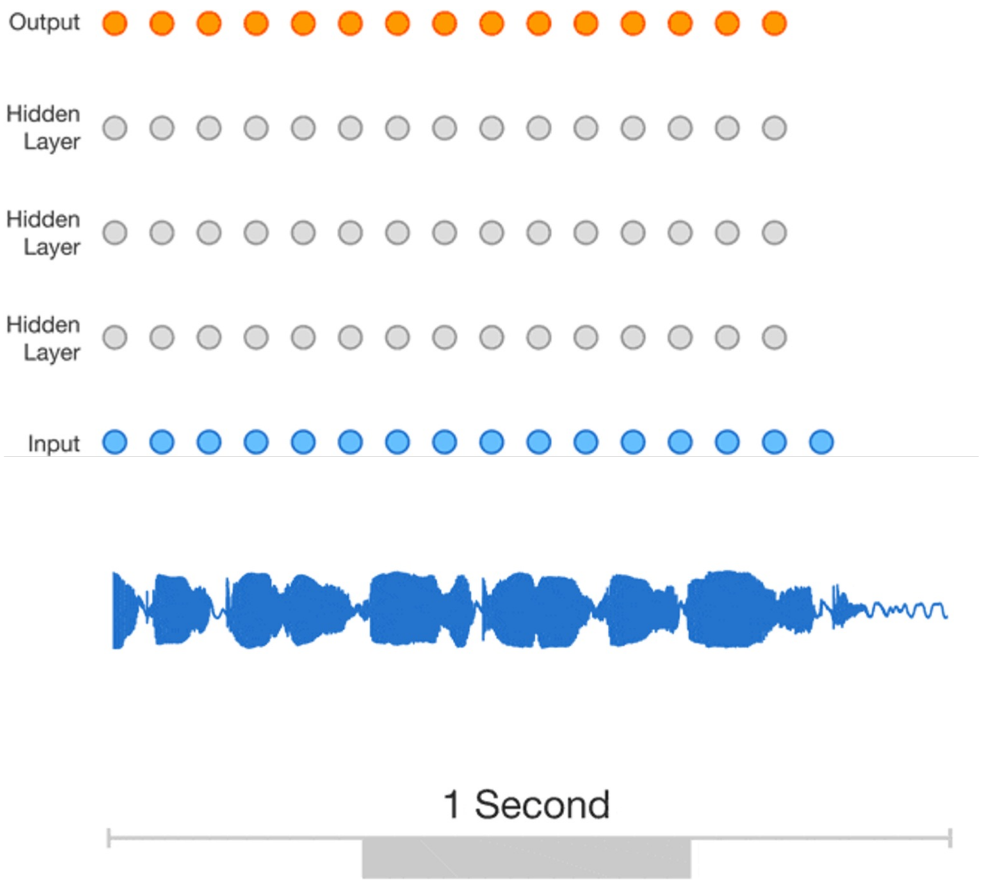


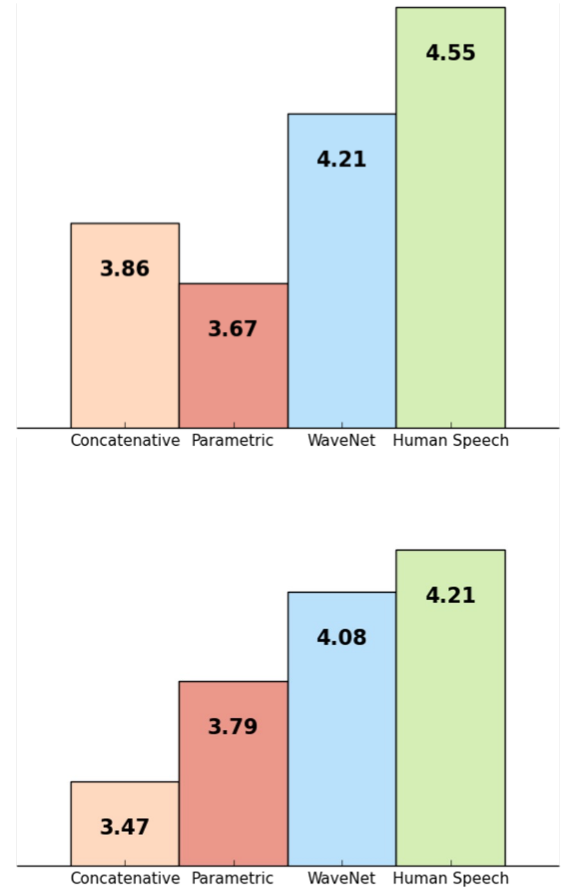
van den Oord, A., et al. “Pixel Recurrent Neural Networks.” *ICML (2016).*

Conditional Pixel CNN

van den Oord, A., et al. “Conditional Pixel CNN.” *NIPS (2016).*

WaveNets

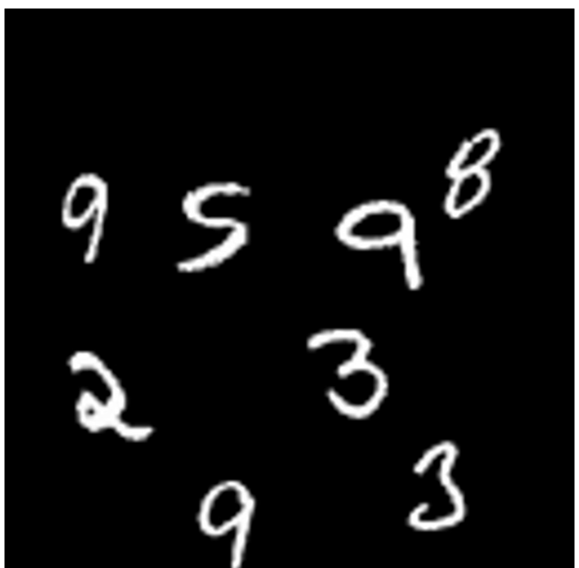




van den Oord, A., et al. “WaveNet: A Generative Model for Raw Audio.” *arxiv (2016).*

Video Pixel Network (VPN)

Model	Test
(Shi et al., 2015)	367.2
(Srivastava et al., 2015a)	341.2
(Brabandere et al., 2016)	285.2
(Patraucean et al., 2015)	179.8
Baseline model	110.1
VPN	**87.6**
Lower Bound	86.3



Kalchbrenner, N., et al. “Video Pixel Networks.” ***ICML (2017)***.

New Architectures