Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 5.2: Alignment and Structured Representations

Louis-Philippe Morency

Objectives of today's class

- Hard Attention – Glimpse model
- Audio Representations and Alignment
  - Connectionist Temporal Classification (CTC)
- Language compositionality and structure
  - Constituency and dependency parsing
- Structured representations
  - Tree-based RNN, Stack LSTM
- VQA and attention models
  - Co-attention, Stacked attention
- Modular neural networks
  - End-to-end learning

Administrative Stuff

Upcoming Schedule

- First project assignment:
  - Proposal presentation (10/1 and 10/3)
  - First project report (Sunday 10/6)
- Midterm project assignment
  - Midterm presentations (11/5 and 11/7)
  - Midterm report (Sunday 11/10)
- Final project assignment
  - Final presentation (12/3 & 12/5)
  - Final report (Sunday 12/8)

Tuesday October 1st – Team Presentations

1	Youtube-8M	Fan Qian, Xue Xia, Yuwei Qiu, Keyi Yu
2	OKVQA	Kaixin Ma, Xiaochuang Han, Meiqi Guo, Zeeshan Ashraf
3	Visual dialogue	Tianwei Yue, Zhihao Zhou, Jiaming Bai, Wenping Wang
4	Argoverse	Nilesh Choubey, Venkat Srinivasan, Tammy Agrawal, Struthi Bannur, Hitesh Arora
5	Embedding fusion in vqa	Chang Gao, Zhiyu Min, Yujia Chen, Yongxin Wang
6	MELD	Aditya Galada, Ritika Mulagalapalli, Roshan Sharma, Siddharth Kannan
7	MIT states	Syed Ashar Javed, Rishi Madhok, Anshuman Majumdar, Talha Siddiqui
8	RefCOCO	Jing Wen, Bereket Frezgiy, Yansen Wang, Parth Shah
9	MOSI	Chengfeng Mao, Michelle Ma, Joohyung Shin

Thursday October 3rd – Team Presentations

1	Esports Twitch	Alex Haig, Wenyan Hu, Vivek Pandit, Longxiang Zhang, Guoxi Zhang
2	TVQA	Victoria Lin, Lucen Zhao, George Xu
3	Audio set	Peter Wu, Muqiao Yang, Zimeng Qiu, Eric Chen & Xinyu Guan
4	MOSEI	Cheng Zhang, Mark Cheung, Yuying Zhu
5	Unsupervised image	Vinayshekhar Bannihatti kumar, Varun Rao, Prakhar Gupta, Mukul Bhutani
6	CLEVR-dialog	Muhammad Shah, Shikib Mehri, Tejas Srinivasan, Vaibhav Kumar
7	Talk the Walk	C R Madhavan, Furqan Khwaja, Harshwardhan Lodha, Anupma Sharan
8	Dialogue image retrieval	Evgeniia Razumovskaia, Ksenia Korovina, Jiaxu Zou
9	Argoverse	Seong Hyeon Park, Gyubok Lee, Minseok Kang, Ashwin Jadhav, Manoj Bhat

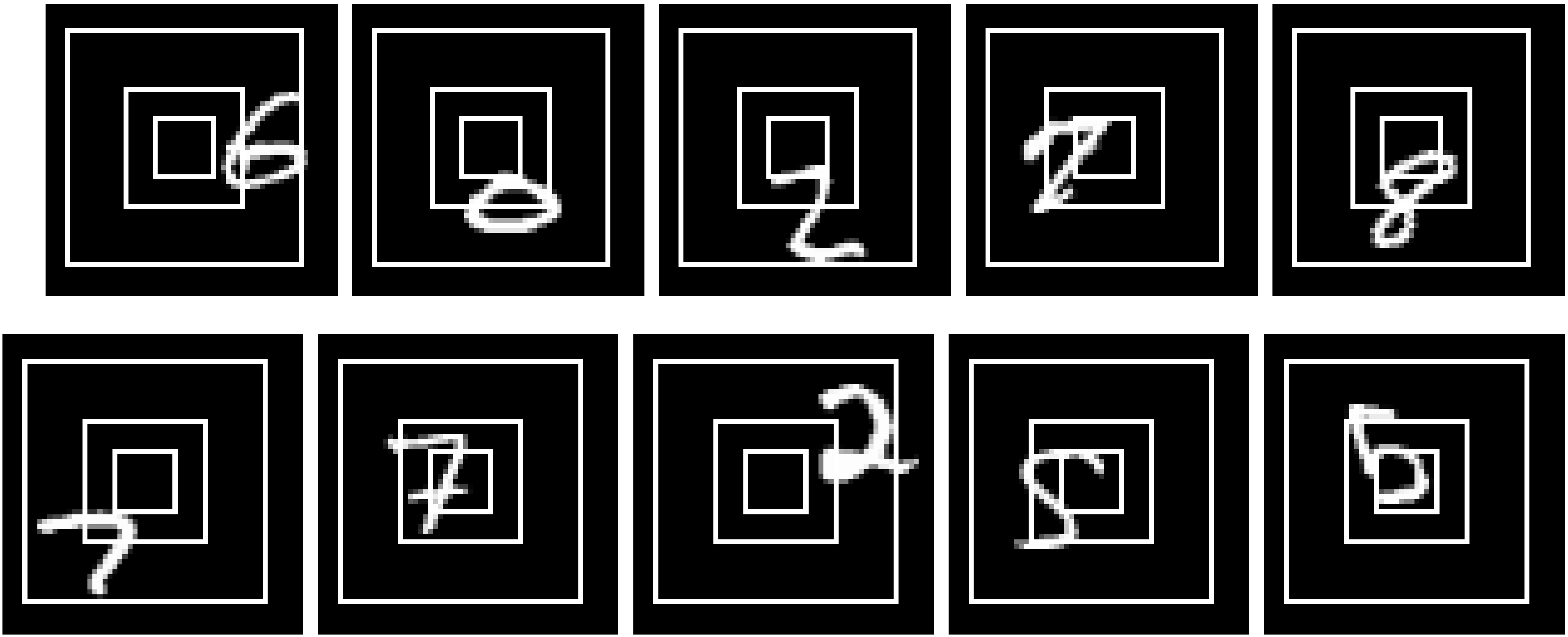
Glimpse Network (Hard Attention)

Hard attention

- Soft attention requires computing a representation for the whole image or sentence
- Hard attention on the other hand forces looking only at one part
- Main motivation was reduced computational cost rather than improved accuracy (although that happens a bit as well)
- **Saccade followed by a glimpse – how human visual system works**

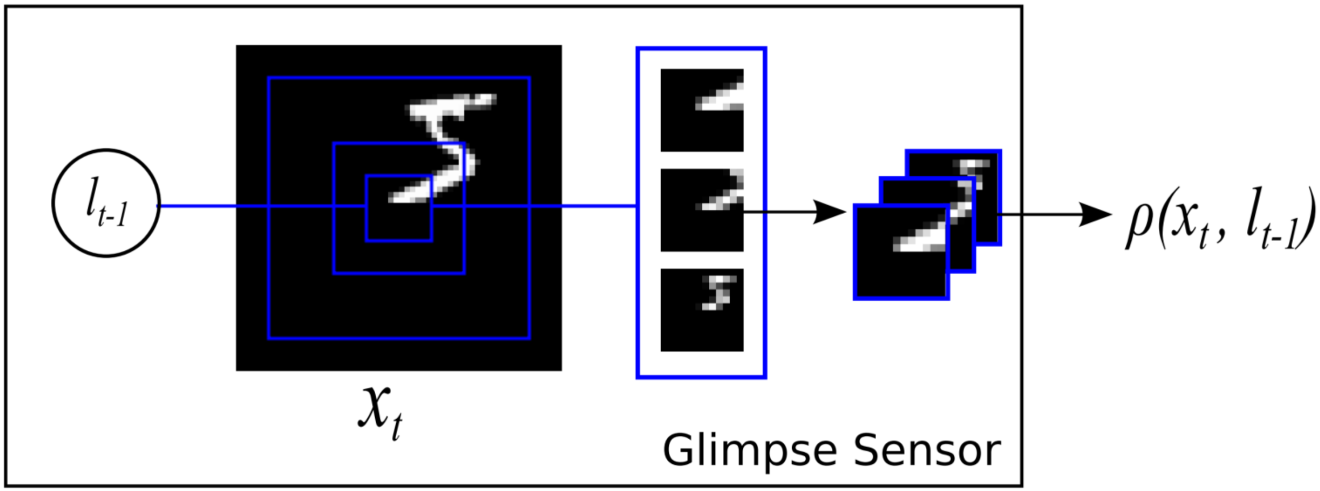
[Recurrent Models of Visual Attention, Mnih, 2014]
[Multiple Object Recognition with Visual Attention, Ba, 2015]

Hard attention examples



Glimpse Sensor

- Looking at a part of an image at different scales

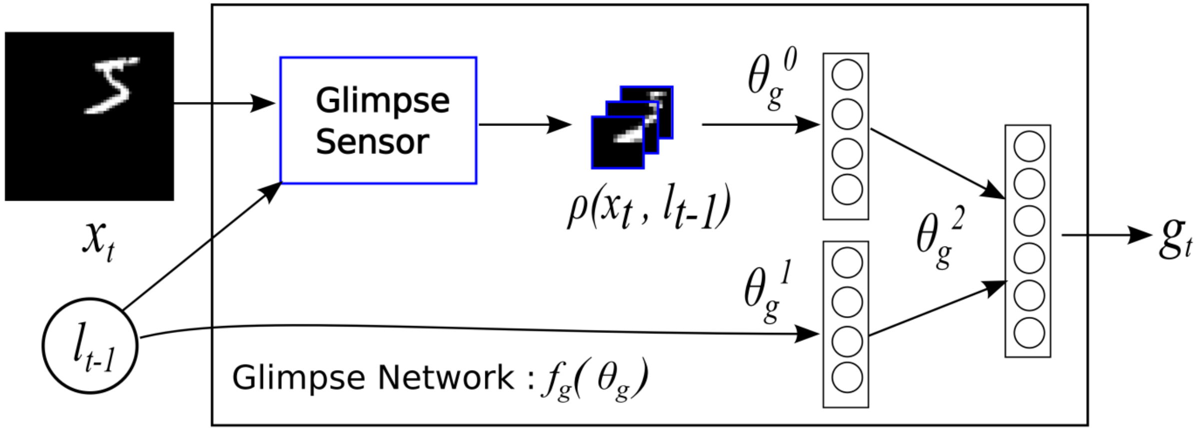


- At a number of different scales combined to a single multichannel image (human retina like representation)
- Given a location l_t output an image summary at that location

[Recurrent Models of Visual Attention, Mnih, 2014]

Glimpse network

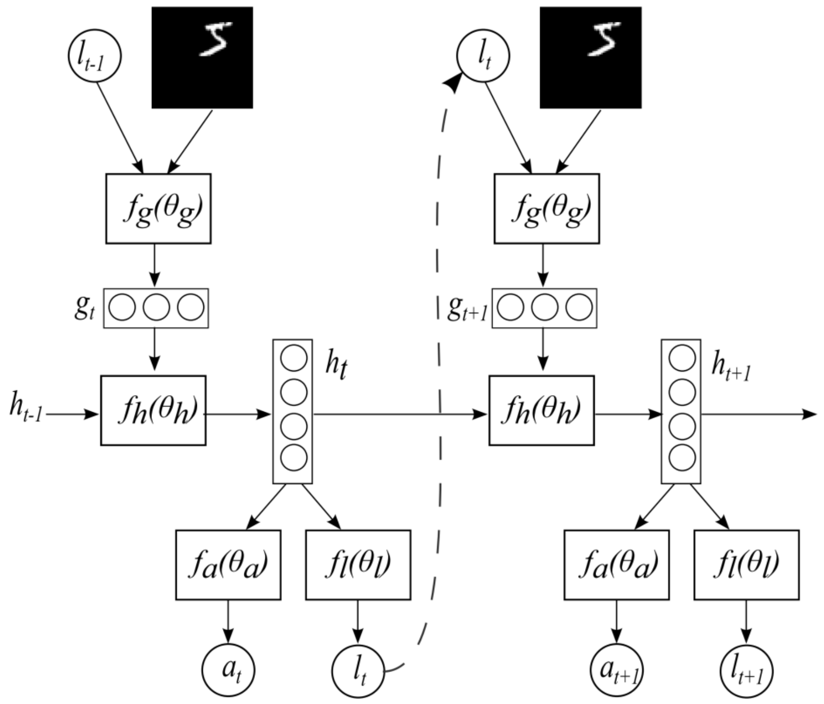
- Combining the Glimpse and the location of the glimpse into a joint network



- The glimpse is followed by a feedforward network (CNN or a DNN)
- The exact formulation of how the location and appearance are combined varies, the important thing is combining **what** and **where**
- Differentiable with respect to glimpse parameters but not the location

Overall Architecture - Emission network

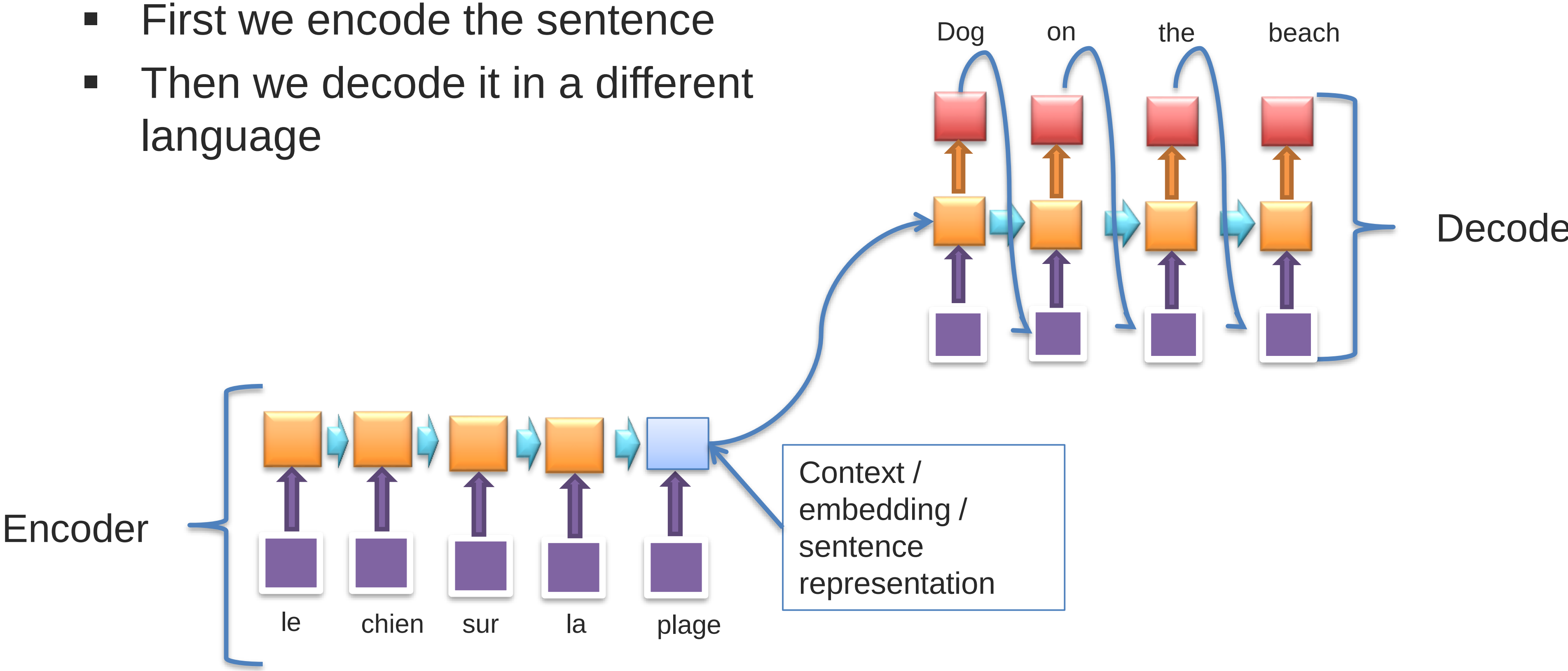
- Given an image a glimpse location l_t, and optionally an action a_t
- Action can be:
  - Some action in a dynamic system – press a button etc.
  - Classification of an object
  - Word output
- This is an RNN with two output gates and a slightly more complex input gate!



Sequence-to-Sequence

Sequence-to-Sequence for Machine Translation

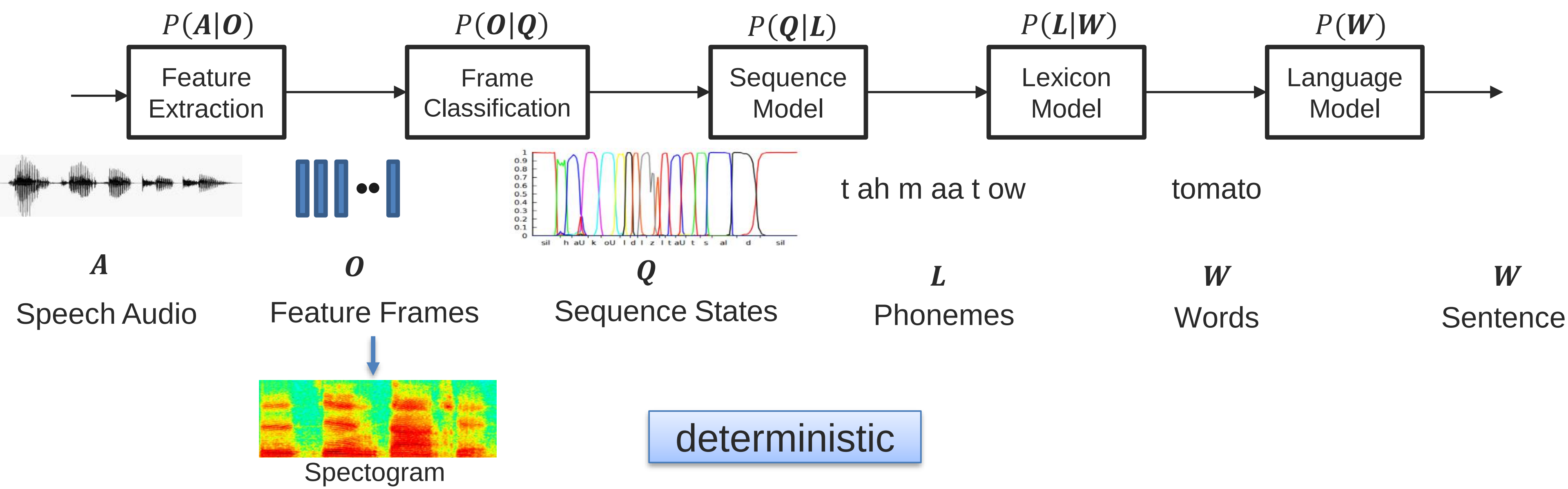
- A quick reminder about encoder decoder frameworks
- First we encode the sentence
- Then we decode it in a different language



Architecture of Speech Recognition

$$\widehat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}} P(\boldsymbol{W}|\boldsymbol{O})$$

$$= \underset{\boldsymbol{W}}{\operatorname{argmax}} P(\boldsymbol{A}|\boldsymbol{O})P(\boldsymbol{O}|\boldsymbol{Q})P(\boldsymbol{Q}|\boldsymbol{L})P(\boldsymbol{L}|\boldsymbol{W})P(\boldsymbol{W})$$

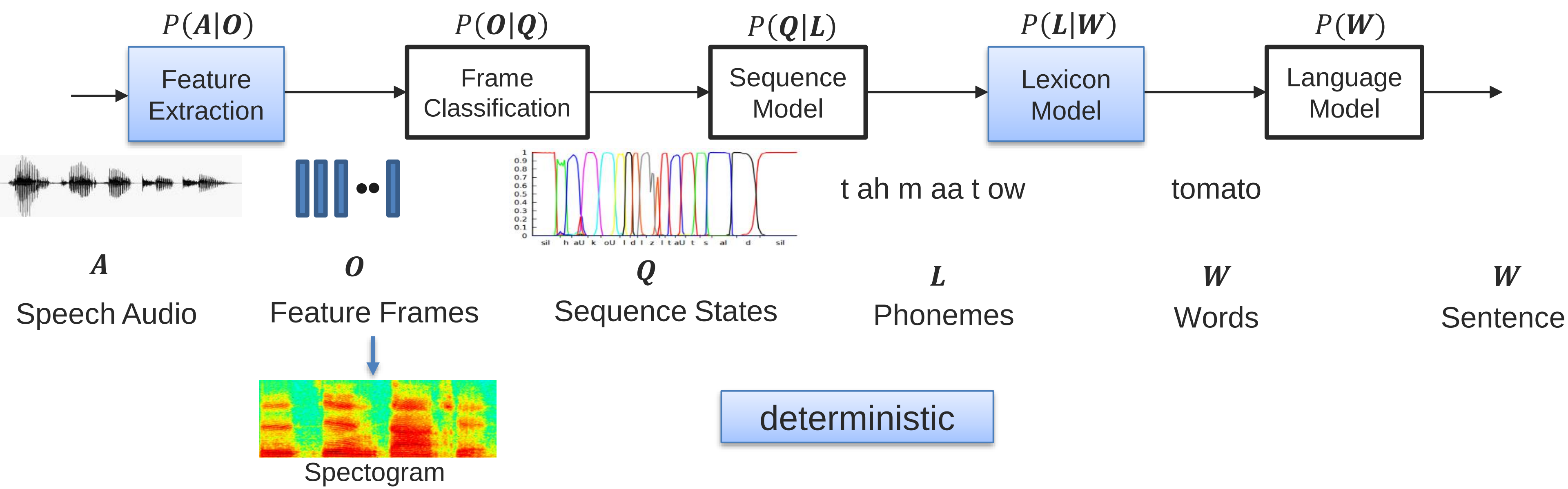


http://slazebni.cs.illinois.edu/spring17/lec26_audio.pdf

Architecture of Speech Recognition

$$\widehat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}}\, P(\boldsymbol{W}|\boldsymbol{O})$$

$$= \underset{\boldsymbol{W}}{\operatorname{argmax}}\, P(\boldsymbol{A}|\boldsymbol{O})P(\boldsymbol{O}|\boldsymbol{Q})P(\boldsymbol{Q}|\boldsymbol{L})P(\boldsymbol{L}|\boldsymbol{W})P(\boldsymbol{W})$$

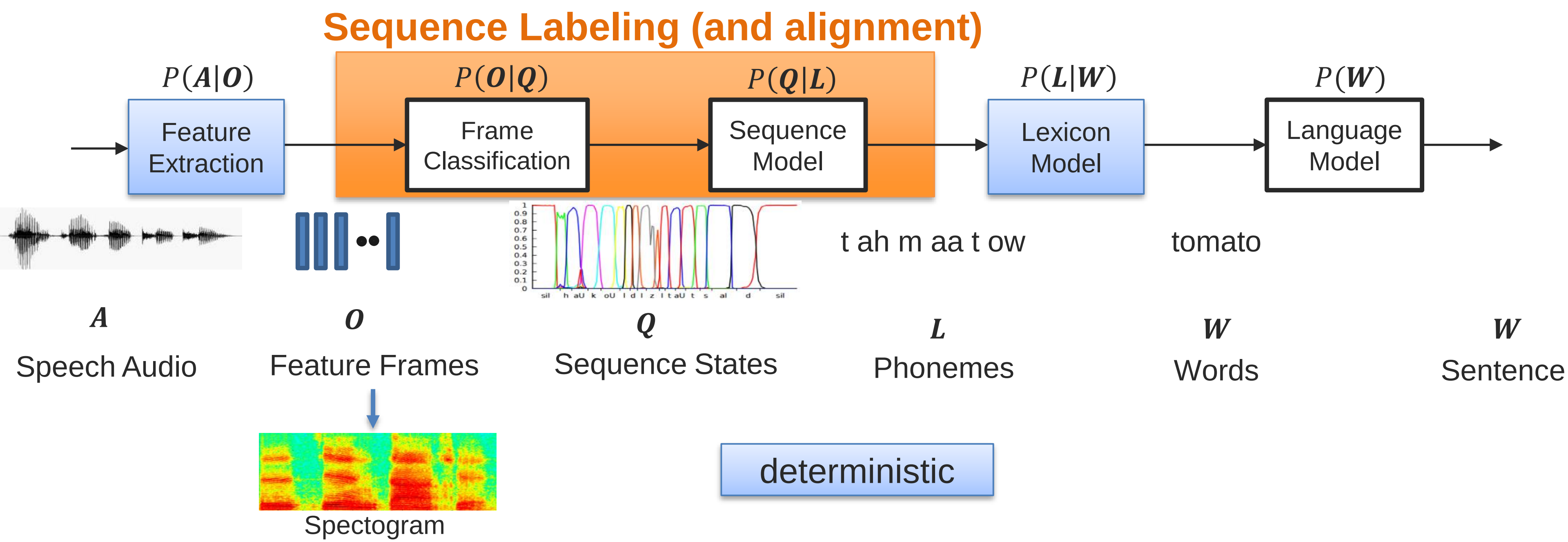


http://slazebni.cs.illinois.edu/spring17/lec26_audio.pdf

Architecture of Speech Recognition

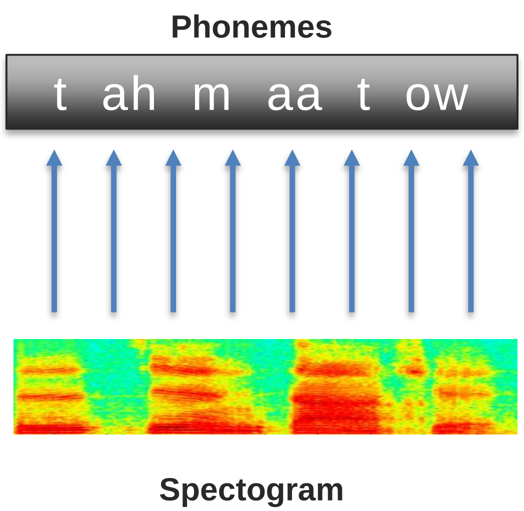
$$\widehat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}}\, P(\boldsymbol{W}|\boldsymbol{O})$$

$$= \underset{\boldsymbol{W}}{\operatorname{argmax}}\, P(\boldsymbol{A}|\boldsymbol{O})P(\boldsymbol{O}|\boldsymbol{Q})P(\boldsymbol{Q}|\boldsymbol{L})P(\boldsymbol{L}|\boldsymbol{W})P(\boldsymbol{W})$$



http://slazebni.cs.illinois.edu/spring17/lec26_audio.pdf

Sequence Labeling (and Alignment)



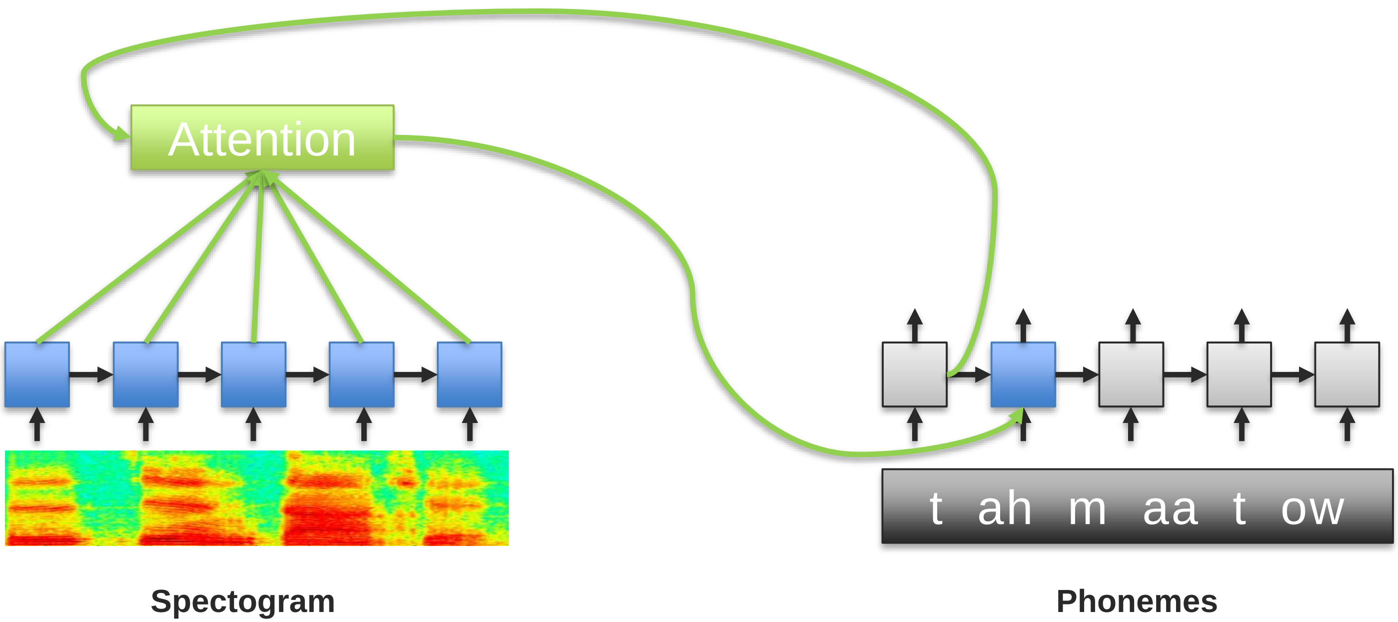
How can we predict the sequence of phoneme labels from the sequence of audio frames?

Option 1: Sequence-to-Sequence (Seq2Seq)

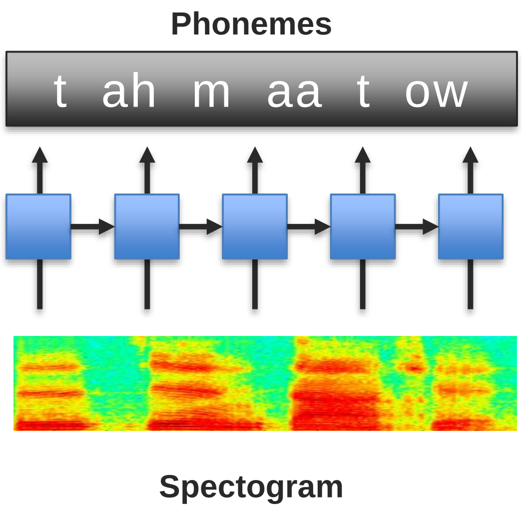
Spectogram

Phonemes

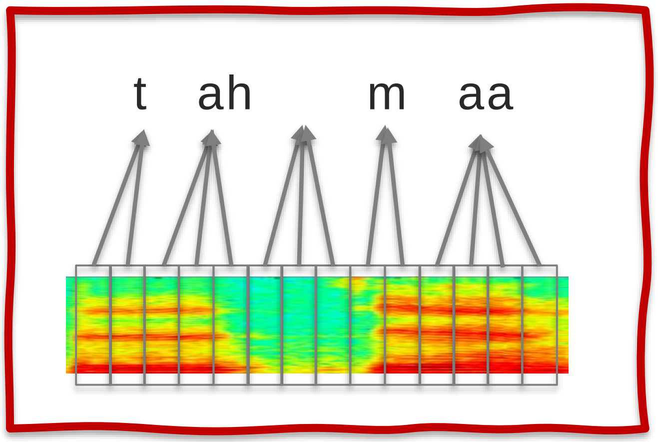
Option 2: Seq2Seq with Attention



Option 3: Sequence Labeling with RNN



Challenge: many-to-1 alignment



What should be the loss function?

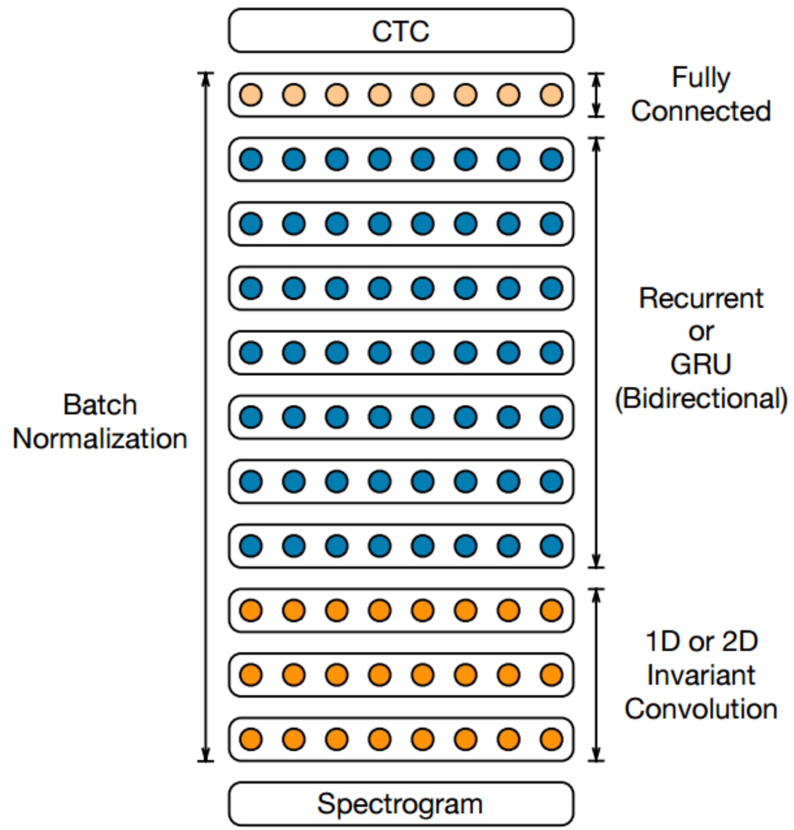
Connectionist Temporal Classification

Connectionist Temporal Classification (CTC)

CTC is used in speech recognition systems that are almost in par with human performances.

Test set	Deep speech 2	Human
WSJ eval'92	3.60	5.03
WSJ eval'93	4.98	8.08
LibriSpeech test-clean	5.33	5.83
LibriSpeech test-other	13.25	12.69

Deep Speech 2



Amodei, Dario, et al. "Deep speech 2: End-to-end speech recognition in english and mandarin." (2015)

Connectionist Temporal Classification (CTC)

Training examples $S = \{(\boldsymbol{x_1}, \boldsymbol{z_1}), \ldots (\boldsymbol{x_N}, \boldsymbol{z_N})\} \in \mathcal{D}_{\mathcal{X}\times\mathcal{Z}}$

$\boldsymbol{x} \in \mathcal{X}$ are spectrogram frames

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_T)$$

$\boldsymbol{z} \in \mathcal{Z}$ are phoneme transcripts

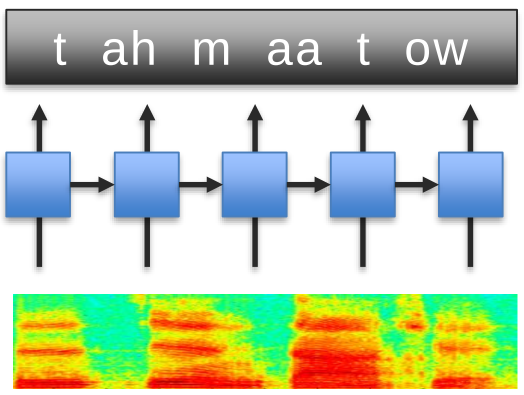
$$\boldsymbol{z} = (z_1, z_2, \ldots, z_U)$$

Not the same length

$U \leq T$

defined over the space of labels L

Phonemes (z)



Spectogram (x)

Goal: train temporal classifier $h : \mathcal{X} \to \mathcal{Z}$

Loss: Negative log likelihood

$$L(S;\theta) = -\sum_{(\boldsymbol{x},\boldsymbol{z})\in S} \ln\big(p_\theta(\boldsymbol{z}|\boldsymbol{x})\big)$$

Connectionist Temporal Classification (CTC)

Rule-based alignment:

1) Remove all blanks
2) Remove repeated labels

$l = \{a\}$	$l = \{bee\}$
`_aaa____`	`bbbeee_ee`
`___aaaa_`	`_bb_ee__e`
`_aaaaaaa`	`__bbbe_e_`

Temporal alignment

③ Predicted labels $\boldsymbol{l}$

$$P(\boldsymbol{l}|\boldsymbol{x}) = \sum_{\boldsymbol{\pi}} P(\boldsymbol{l}|\boldsymbol{\pi}) P(\boldsymbol{\pi}|\boldsymbol{x})$$

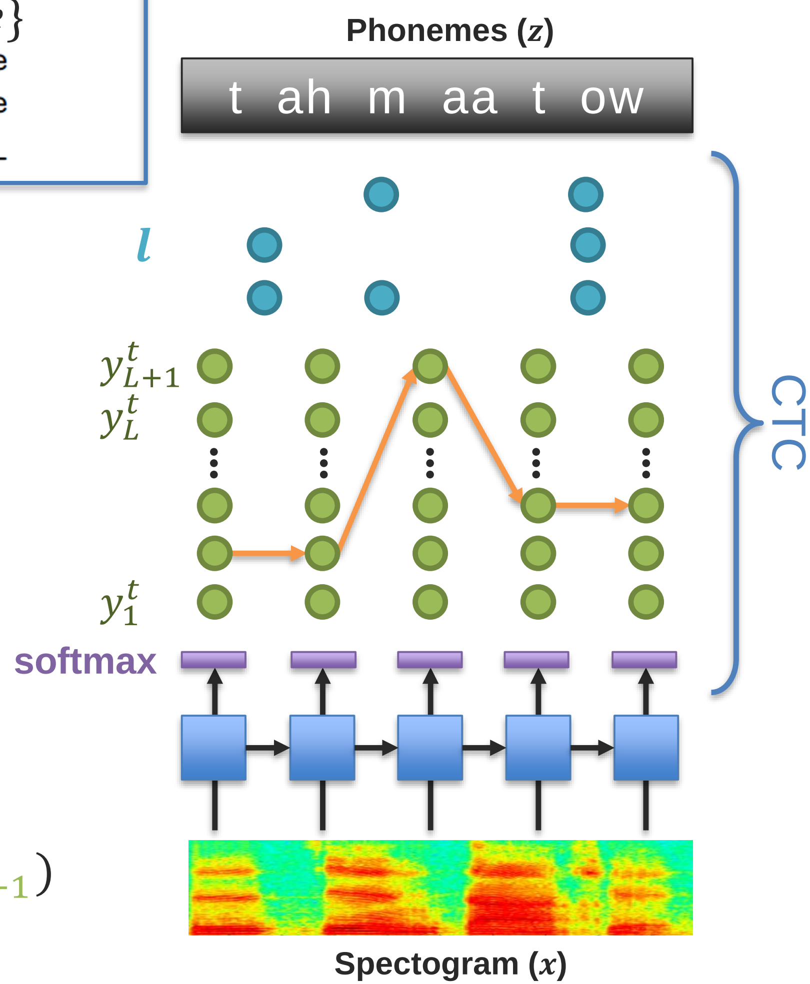
② Path $\boldsymbol{\pi}$ over the activations:

$$\mathrm{P}(\boldsymbol{\pi}|\boldsymbol{x}) = \prod_{t=1}^{T} y^t_{\boldsymbol{\pi}_t}, \forall \boldsymbol{\pi} \in L'^T$$

① Output activations (distribution):

$$\boldsymbol{y} = f_\theta(\boldsymbol{x}), \text{ where } \boldsymbol{y}^t = (y^t_1, y^t_2, \dots, y^t_L, y^t_{L+1})$$

for 'blank' or no label



Connectionist Temporal Classification (CTC)

4 Most probable sequence labels

$$\hat{z} = h(x) = \arg\max_{l \in L^T} P(\boldsymbol{l}|\boldsymbol{x})$$

3 Predicted labels $\boldsymbol{l}$

$$P(\boldsymbol{l}|\boldsymbol{x}) = \sum_{\boldsymbol{\pi}} P(\boldsymbol{l}|\boldsymbol{\pi}) P(\boldsymbol{\pi}|\boldsymbol{x})$$

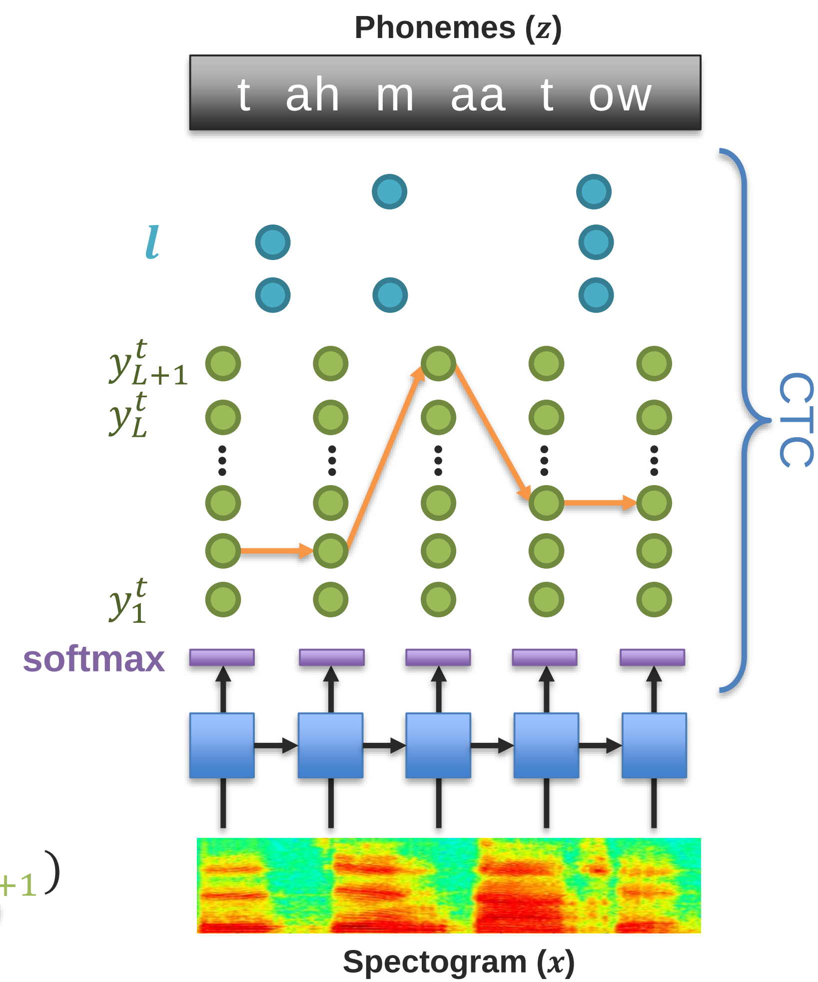
2 Path $\boldsymbol{\pi}$ over the activations:

$$\mathrm{P}(\boldsymbol{\pi}|\boldsymbol{x}) = \prod_{t=1}^{T} y^t_{\boldsymbol{\pi}_t}, \forall \boldsymbol{\pi} \in L'^T$$

1 Output activations (distribution):

$$\boldsymbol{y} = f_\theta(\boldsymbol{x}), \text{ where } \boldsymbol{y}^t = (y^t_1, y^t_2, \dots, y^t_L, y^t_{L+1})$$

for 'blank' or no label



CTC Optimization

4 Most probable sequence labels

$$z^* = h(x) = \arg\max_{l \in L^T} P(\boldsymbol{l}|\boldsymbol{x})$$

Option 1: Select most probable path $\boldsymbol{\pi}$

$$\pi^* = \arg\max_{\pi} P(\boldsymbol{\pi}|\boldsymbol{x})$$

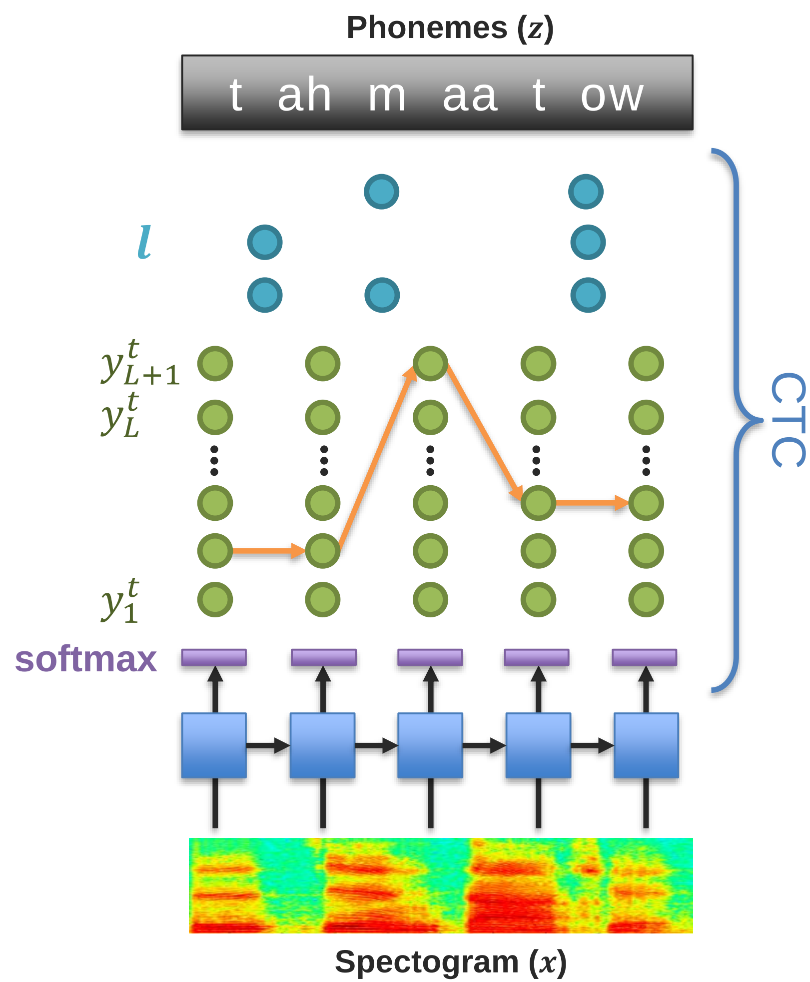
Get most probable labels z^* directly from π^*

Option 2: Solve using dynamic programming

Forward-backward algorithm

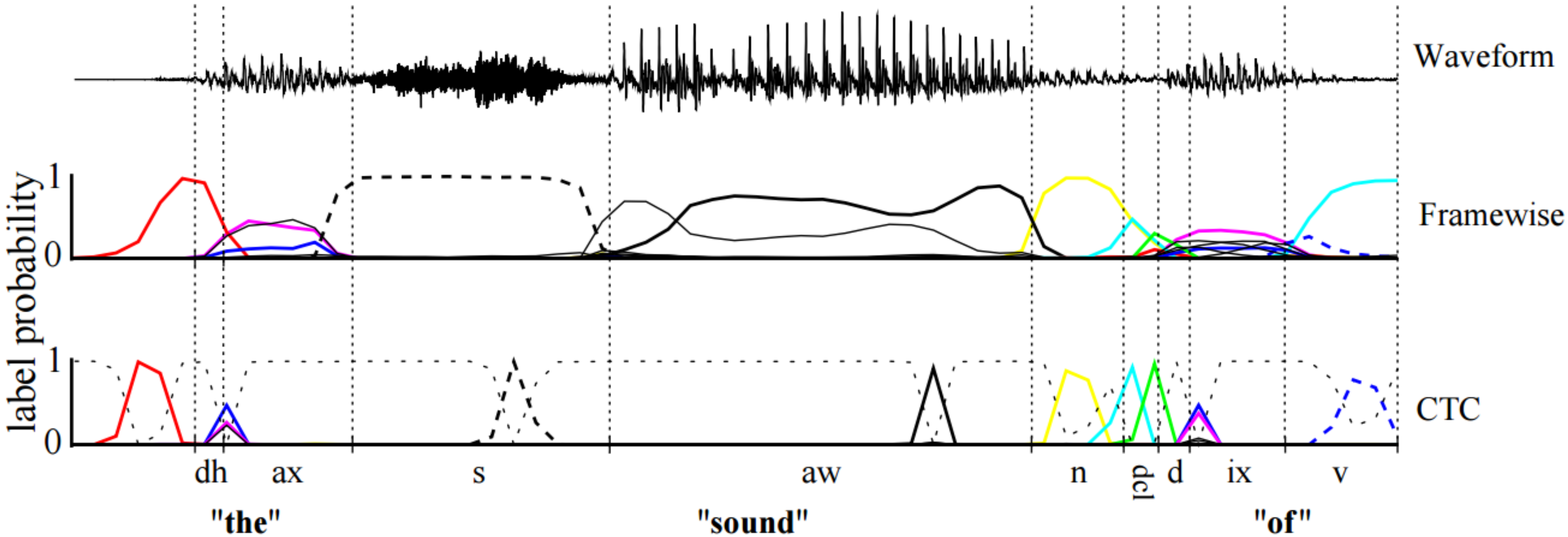
- Forward variables α
- Backward variables β

$$P(l|x) = \sum_{t=1}^{T} \sum_{s=1}^{|l|} \frac{\alpha_t(s)\beta_t(s)}{y_{l_s}^t}$$



Visualizing CTC Predictions

“Framewise” modeling: Learned using phoneme segmentation (vertical lines)

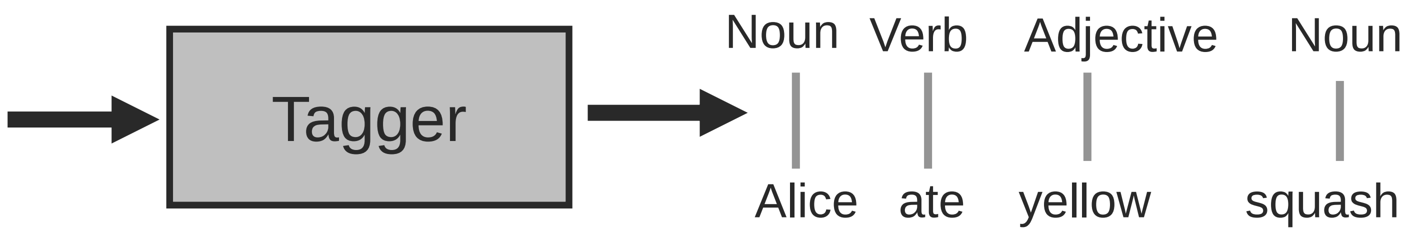


Why are CTC predictions so “peaky”?

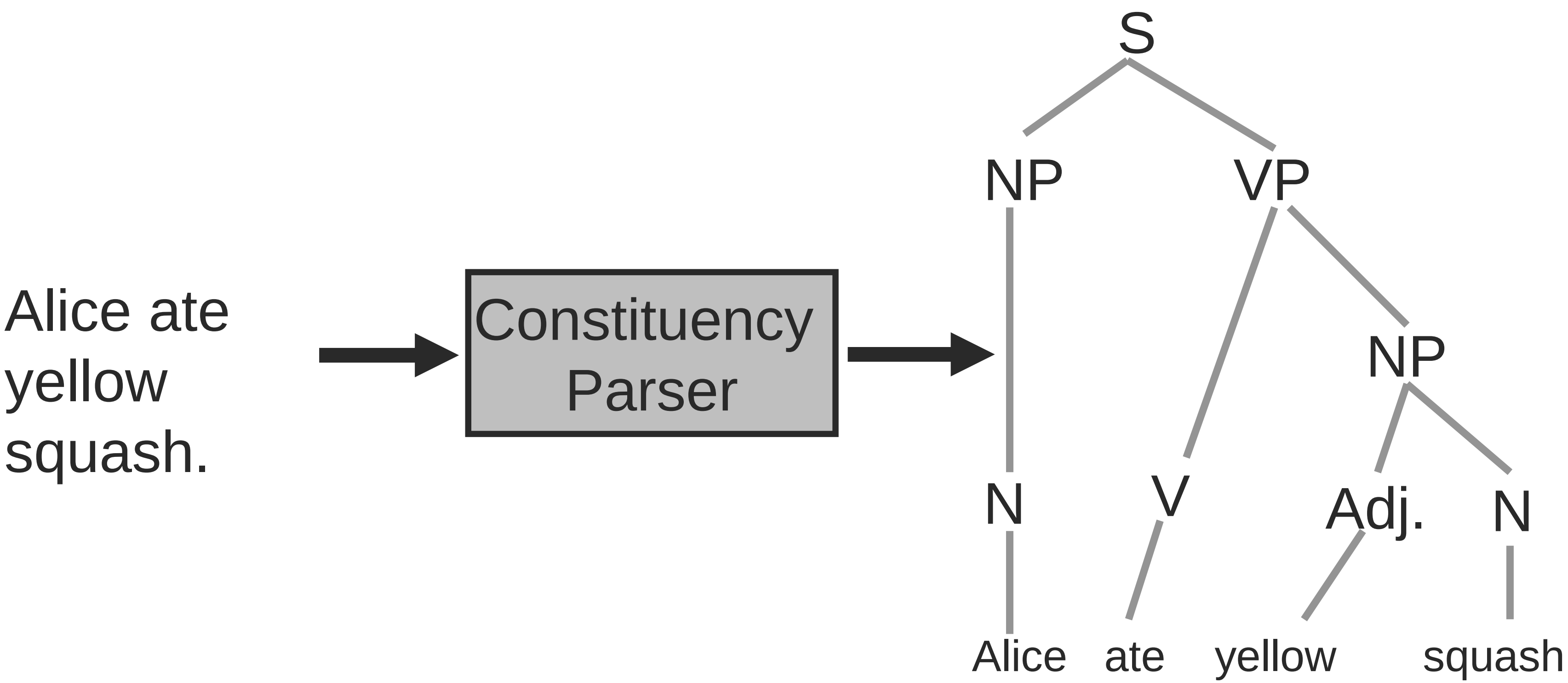
Language Syntax

Part-of-Speech Tagging

Alice ate yellow squash.



Phrase Structure Tree (Constituency Parsing)



The Importance of Parsing

What does “fake” modify?

In the hotel fake property was sold to tourists.

What does “In the hotel” modify?

Phrase Chunking

- Find all non-recursive noun phrases (NPs) and verb phrases (VPs) in a sentence.
  - [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].
  - [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only # 1.8 billion] [PP in] [NP September]

Language Ambiguity

- I saw her duck

vs.

Language Ambiguity

- I saw her duck with a telescope

Language Ambiguity

- I saw her duck with a telescope

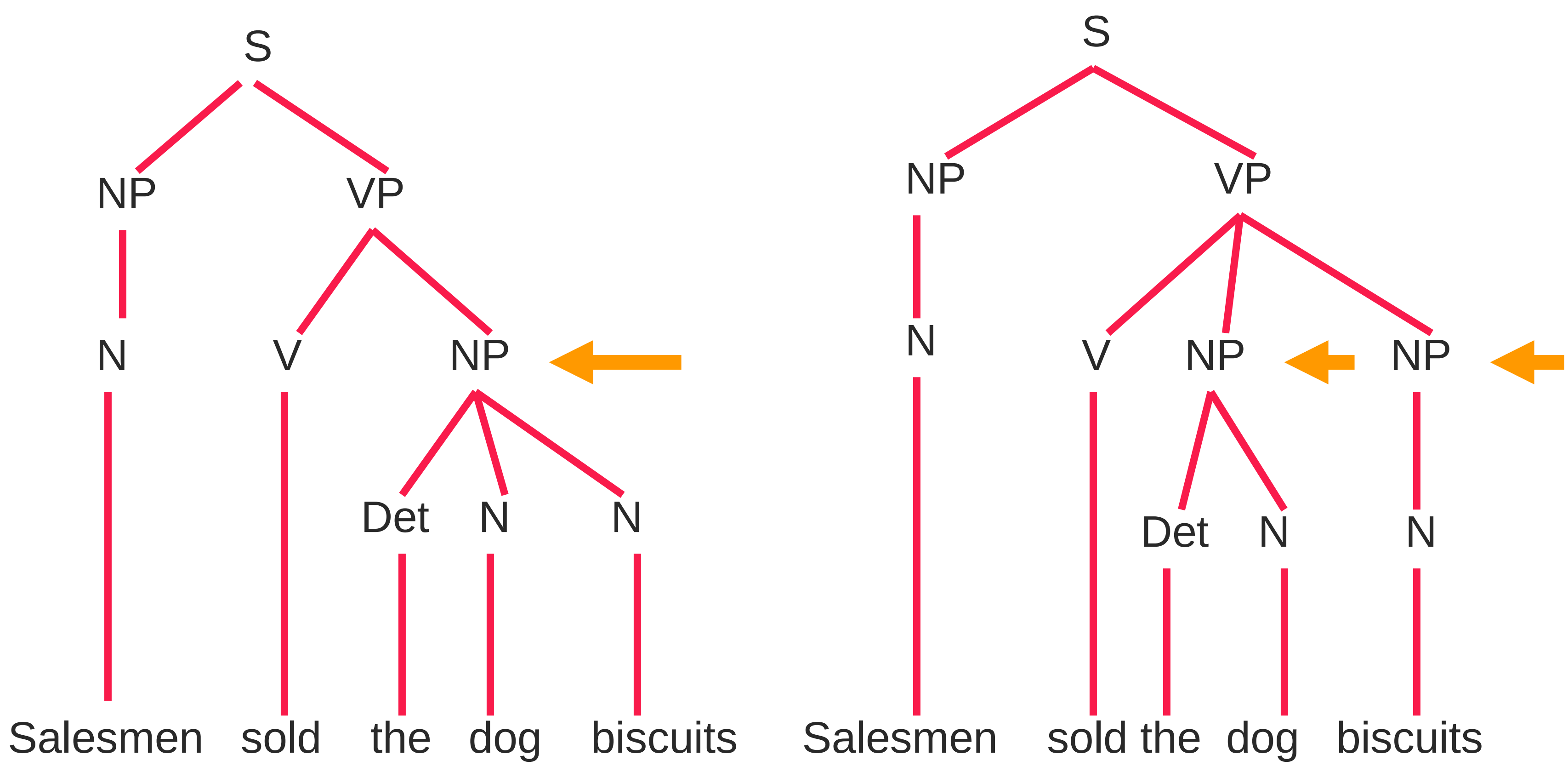
Language Ambiguity

- I saw her duck with a telescope

Language Ambiguity

- I saw her duck with a telescope

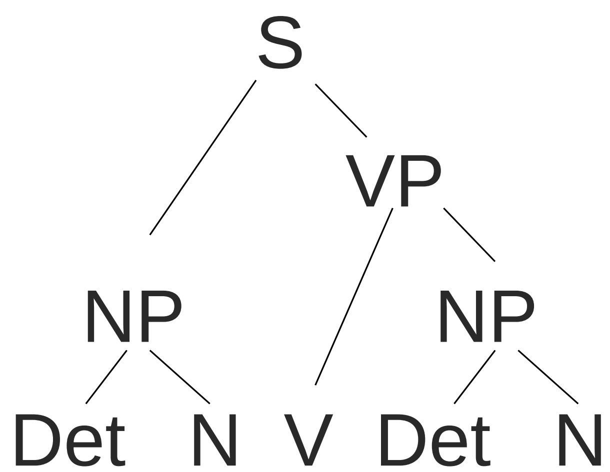
Language Ambiguity



Language Syntax – Examples

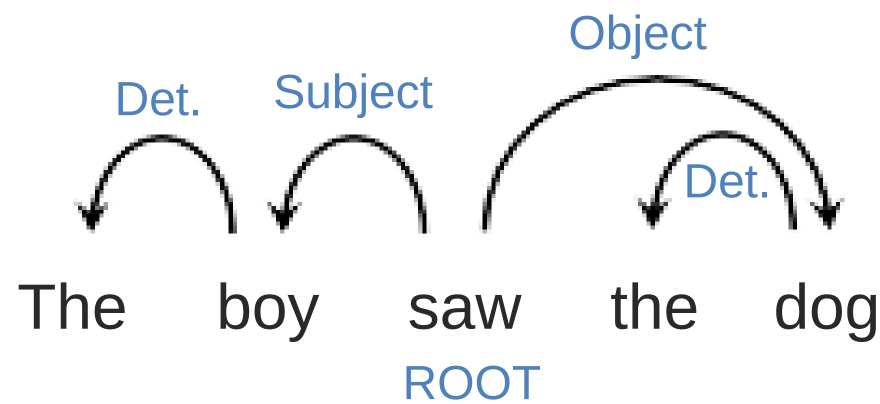
Det	Noun	Verb	Det	Noun	Prep	Det	Noun
The	boy	saw	the	dog	in	the	park

Part of Speech tagging



The boy saw the dog

Constituency Parsing

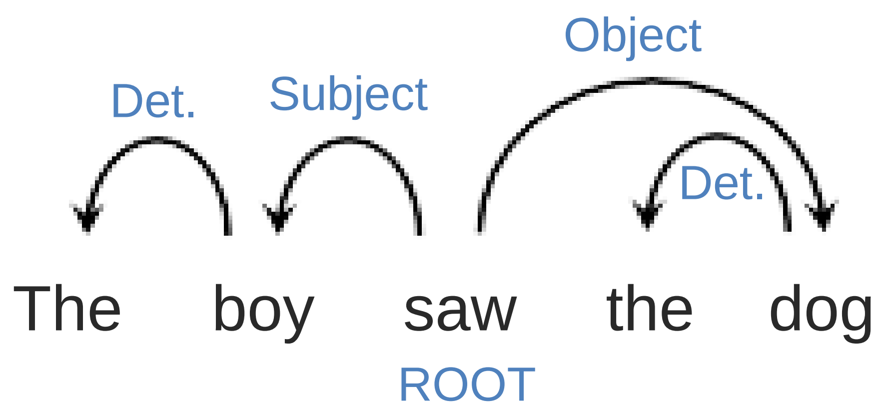


Dependency Parsing

Dependency Syntax

Main idea: Syntactic structure consists of *lexical items*, linked by binary asymmetric relations called *dependencies*

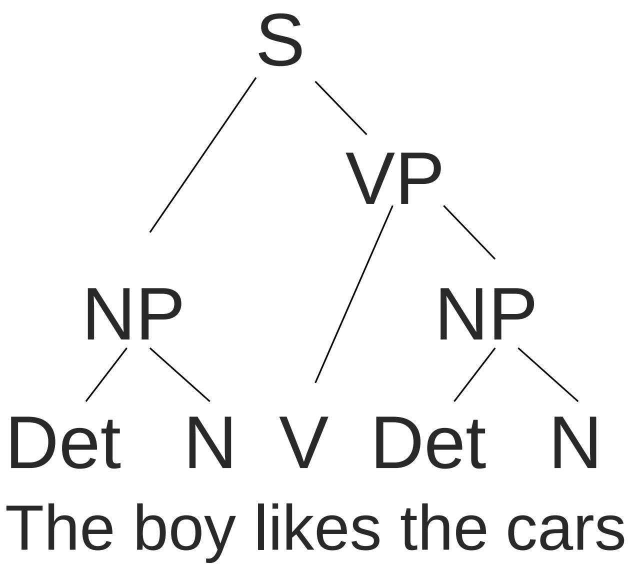
- Easier to convert to predicate-argument structure
- You can try to convert one representation into another
  - But, in general, these formalisms are not equivalent



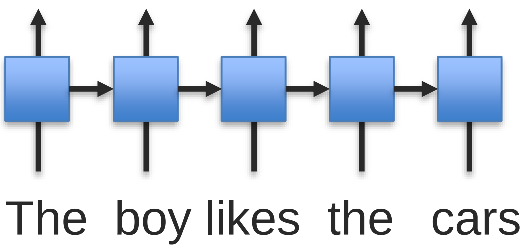
How to take advantage of syntax when modeling language with neural networks?

Recursive Neural Network

How to Model Syntax with RNNs?

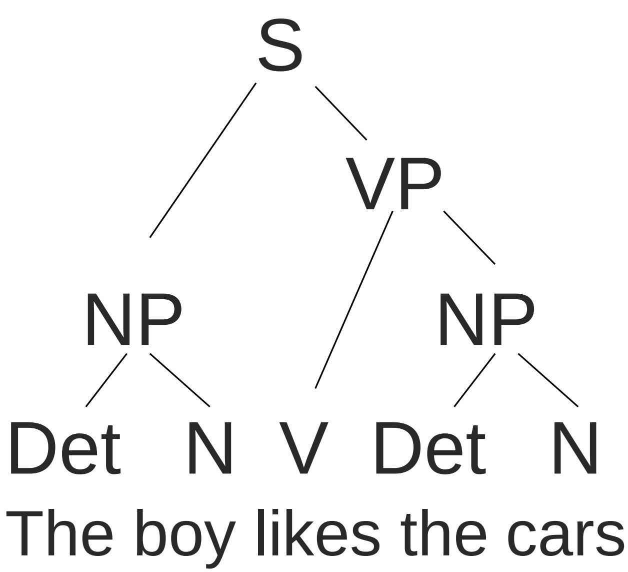


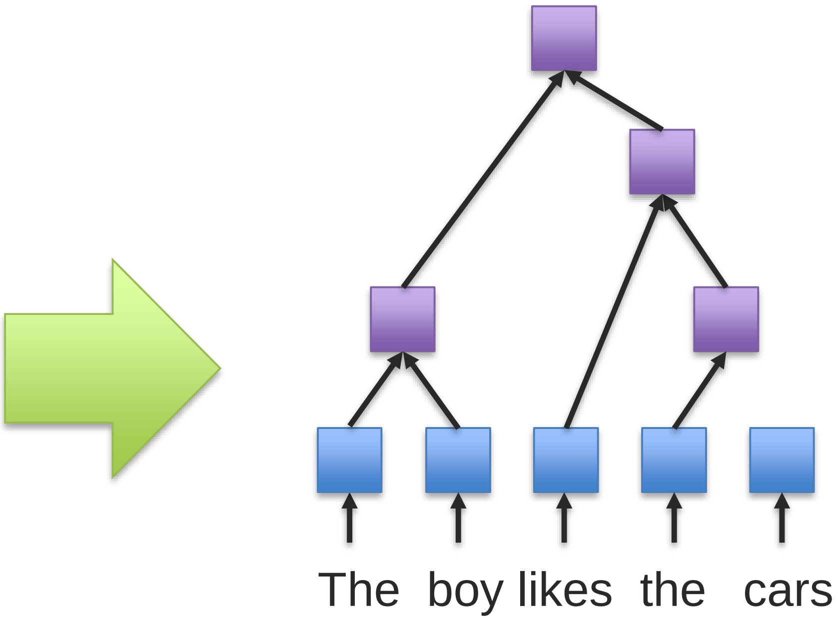
?



We could use Part-of-Speech tags.

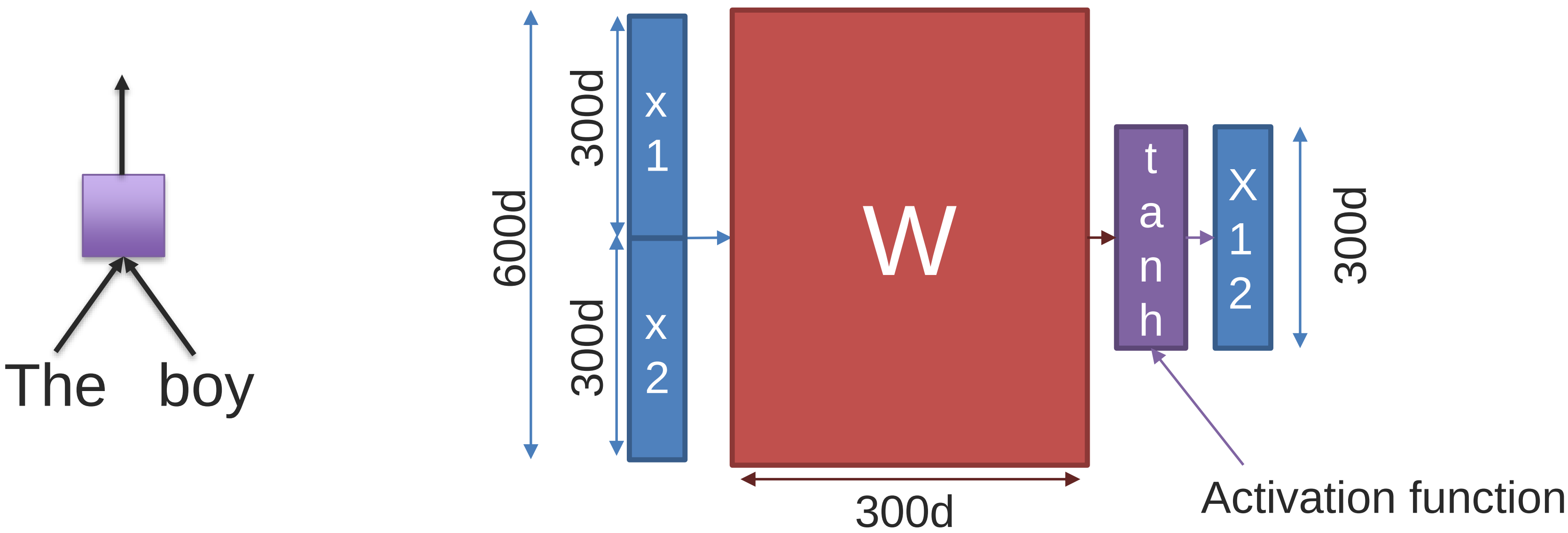
Tree-based RNNs (or Recursive Neural Network)



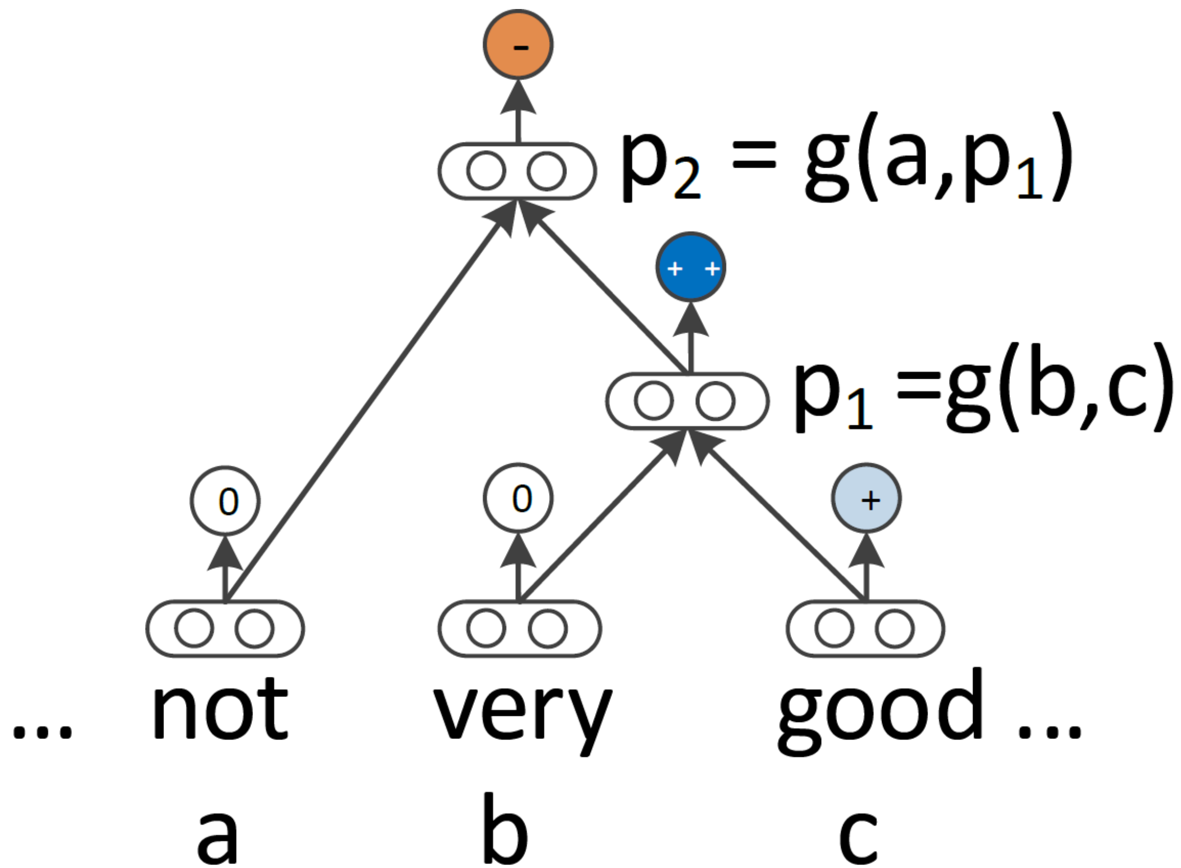


Recursive Neural Unit

- Pair-wise combination of two input features



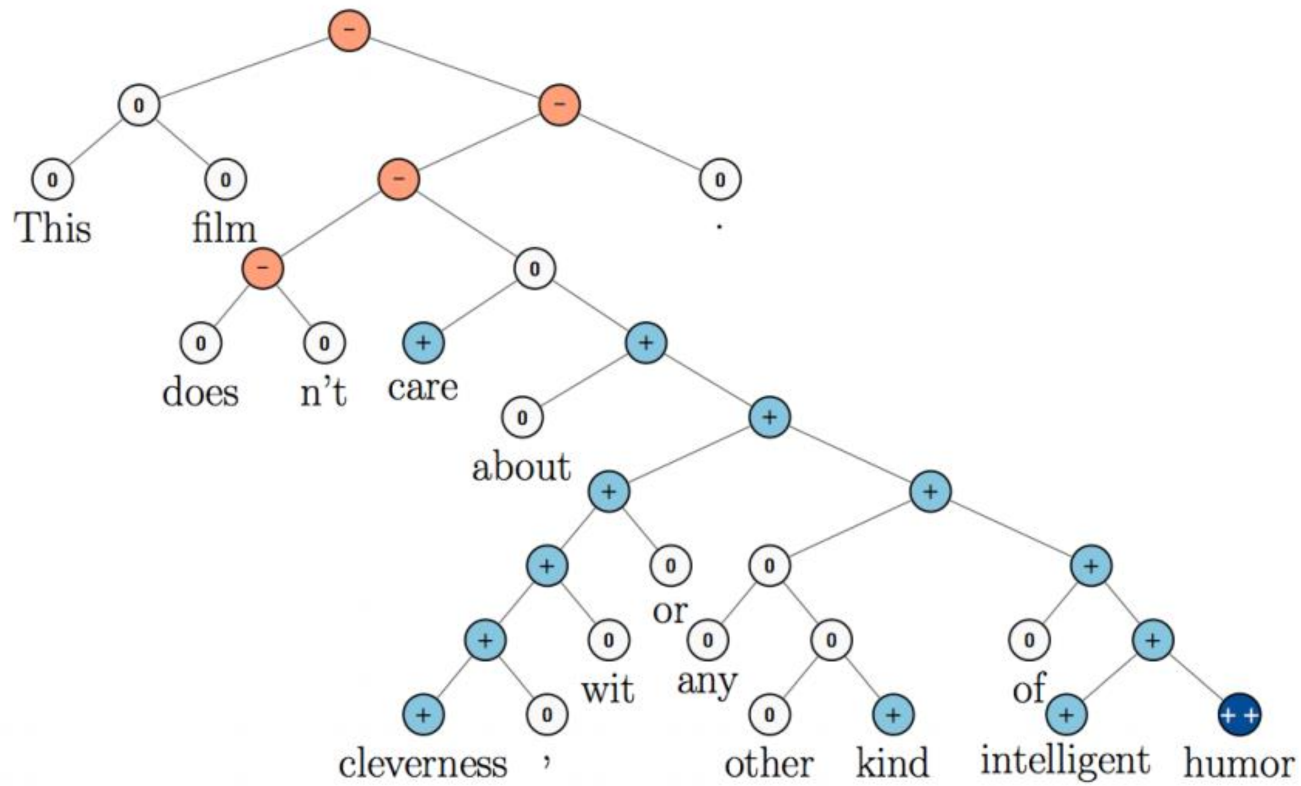
Recursive Neural Network for Sentiment Analysis



Socher et al., Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, EMNLP 2013

Recursive Neural Network for Sentiment Analysis

Classification of a sentence using tree-based compositionality of words



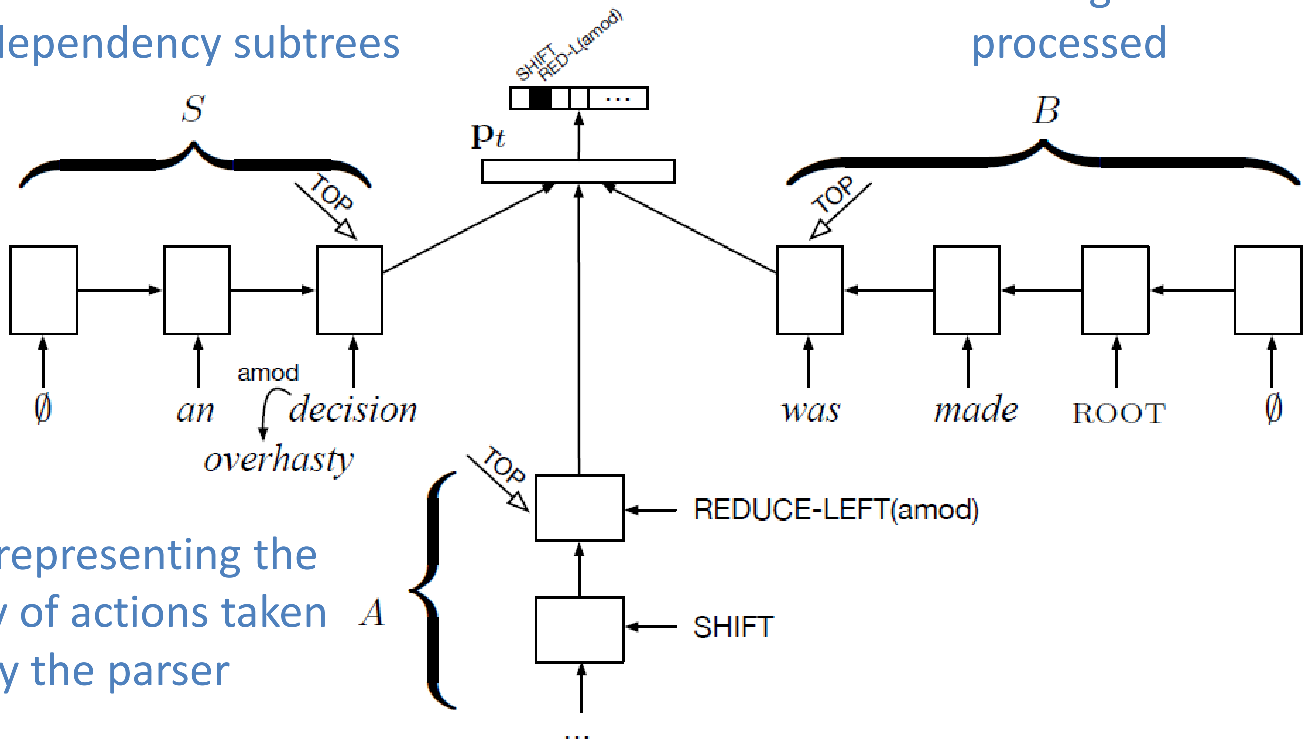
Demo: http://nlp.stanford.edu/sentiment/

Socher et al., Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank, EMNLP 2013

Stack LSTM

stack of partially constructed dependency subtrees

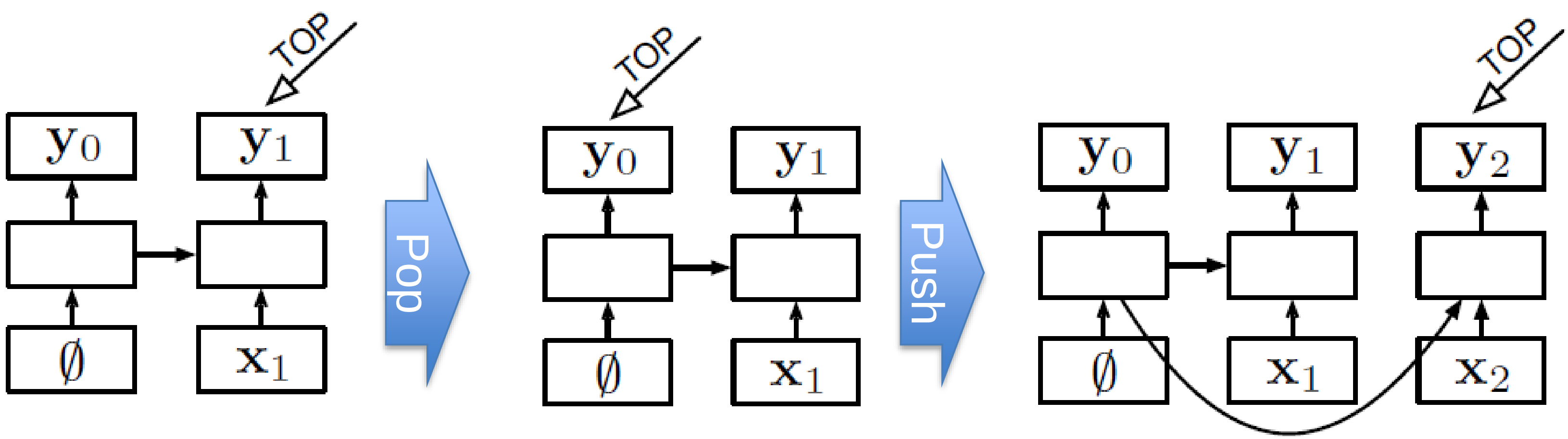
buffer of words remaining to be processed



stack representing the history of actions taken by the parser

Dyer et al., Transition-Based Dependency Parsing with Stack Long Short-Term Memory, 2015

Stack LSTM



Dyer et al., Transition-Based Dependency Parsing with Stack Long Short-Term Memory, 2015

Visual Question Answering And Attention Models

Visual Question Answering

Question

Is the skateboard airborne?

Image

Answer

yes

How can we use attention?

VQA and Attention

Question

Is the skateboard airborne?

Language can
be used to
attend the image

Image

Answer

yes

VQA and Attention

Question

Is the skateboard airborne?

Image could also be used to attend the text

Image

Answer

yes

Co-attention

Question

Is the skateboard airborne?

Image

Or do both!

Answer

yes

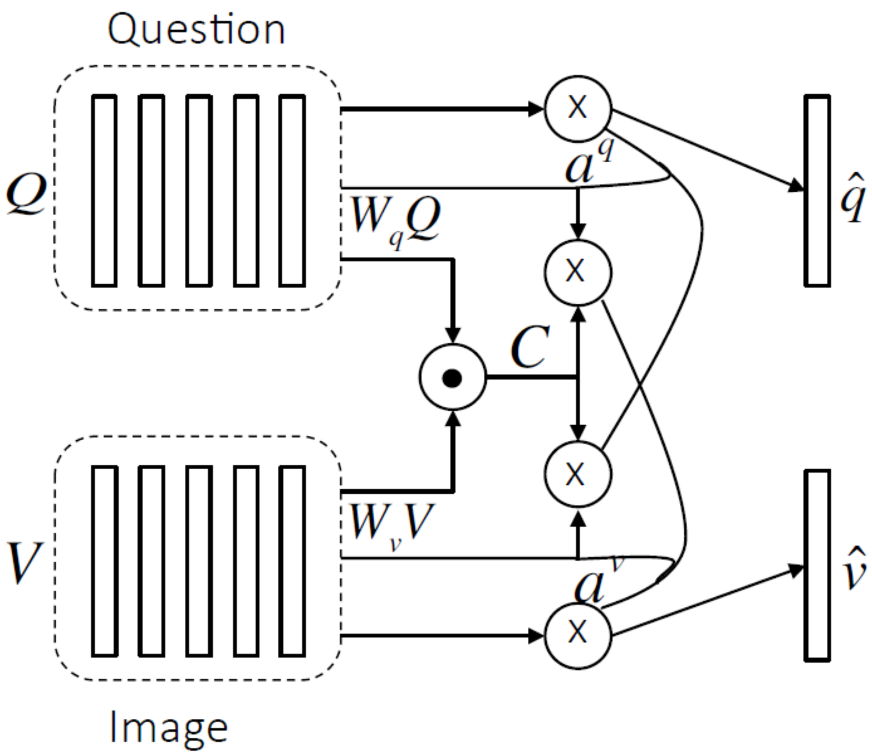
Lu et al., Hierarchical Question-Image Co-Attention for Visual Question Answering, NIPS 2016

Co-attention

Question

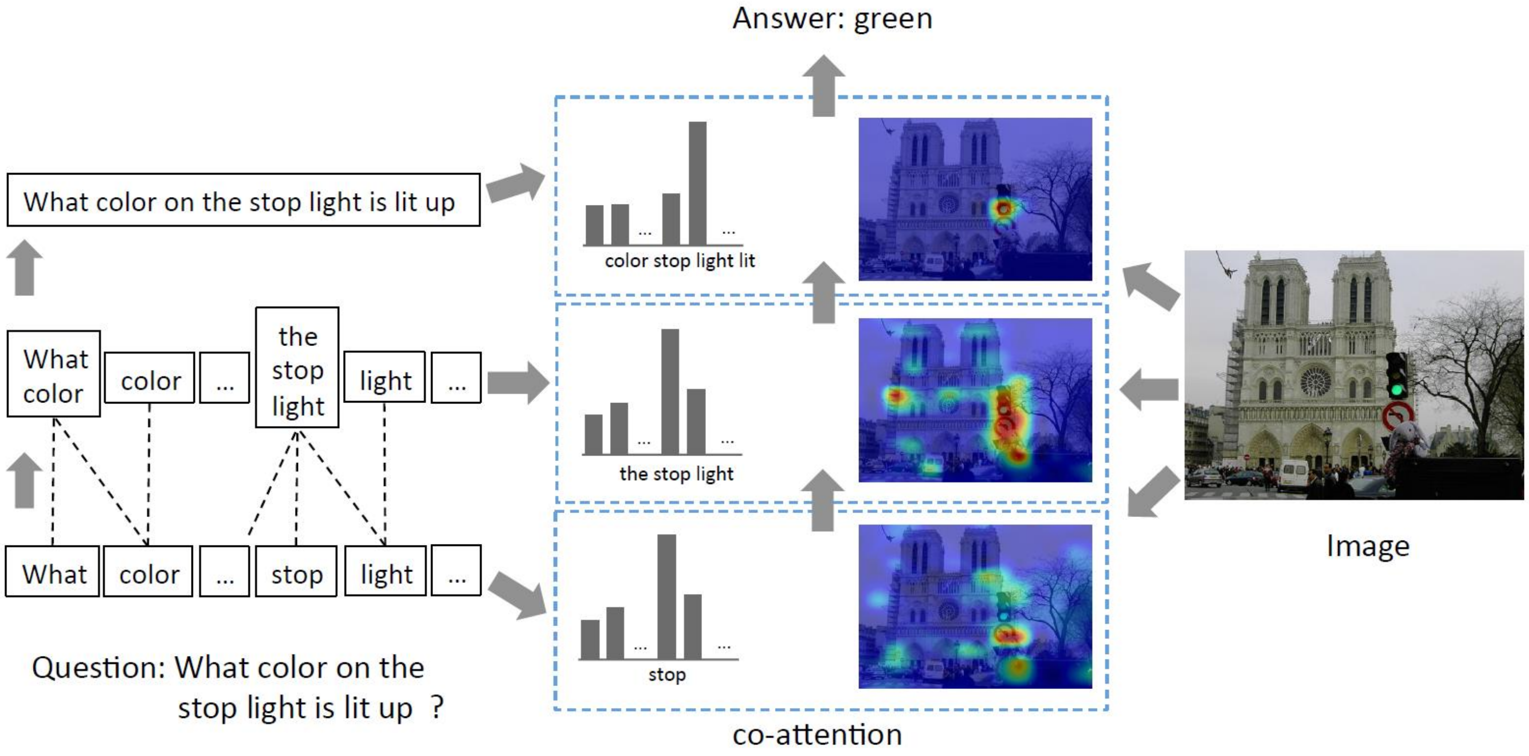
Is the skateboard airborne?

Image



Lu et al., Hierarchical Question-Image Co-Attention for Visual Question Answering, NIPS 2016

Hierarchical Co-attention



Lu et al., Hierarchical Question-Image Co-Attention for Visual Question Answering, NIPS 2016

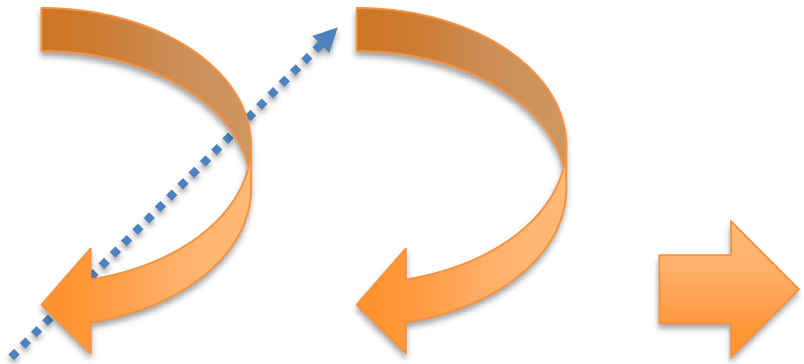
Stacked Attentions

Question

What are sitting in the basket on a bicycle?

Image

Attention 1 Attention 2



Answer

dogs

Attention 1 Attention 2

Yang et al., Stacked Attention Networks for Image Question Answering, CVPR 2016

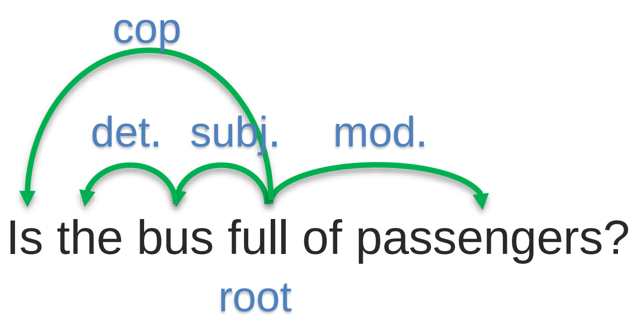
Other Attention-based Models for VQA

- Bottom-up and top-down attention for image captioning and visual question answering, CVPR 2018
  - Adds the idea of object-based representations
- Bilinear Attention Pooling, NIPS 2018
  - Extend low-rank bilinear pooling to multimodal
- Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering, IEEE TNNLS, 2018

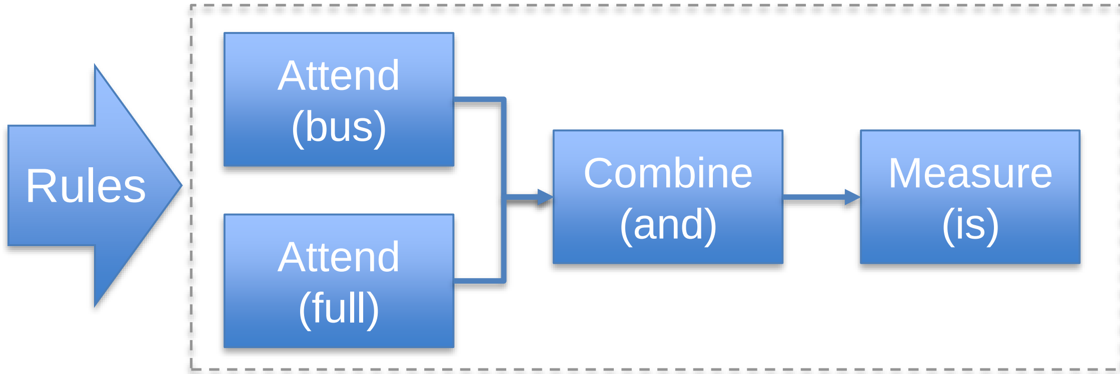
But how to take advantage of language syntax?

Neural Module Networks

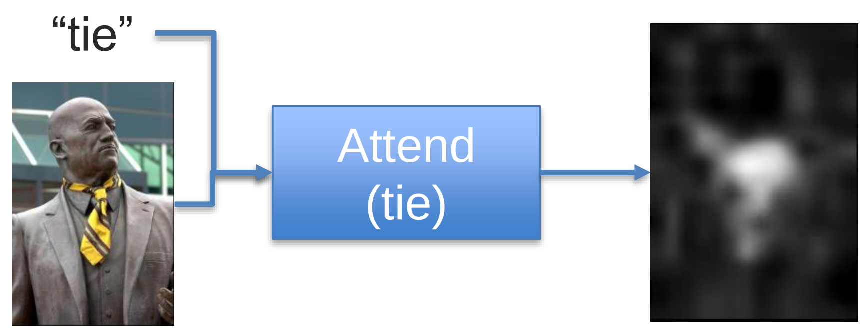
Neural Module Network



Computation layout



Each module work on the attention map(s):

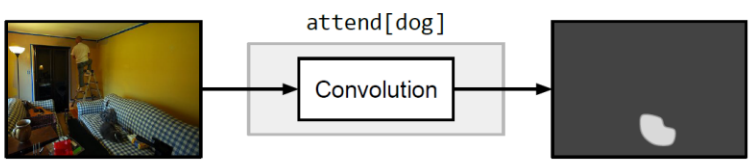


Andreas et al., Deep Compositional Question Answering with Neural Module Networks, 2016

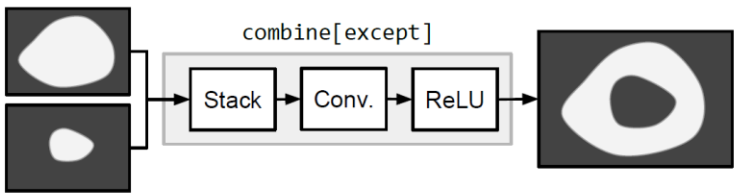
Predefined Set of Modules

1) Analyze the image:

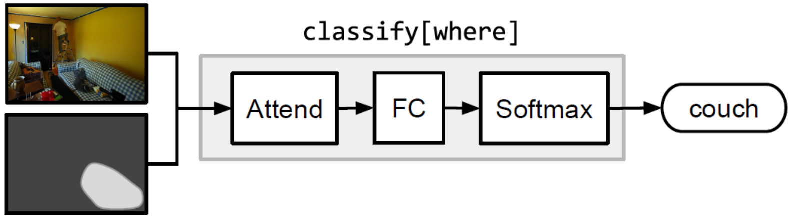
$\texttt{attend} : \mathit{Image} \to \mathit{Attention}$



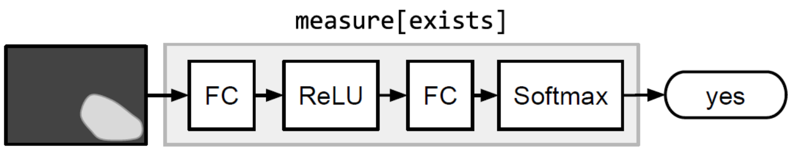
$\texttt{combine} : \mathit{Attention} \times \mathit{Attention} \to \mathit{Attention}$



2) Make a prediction



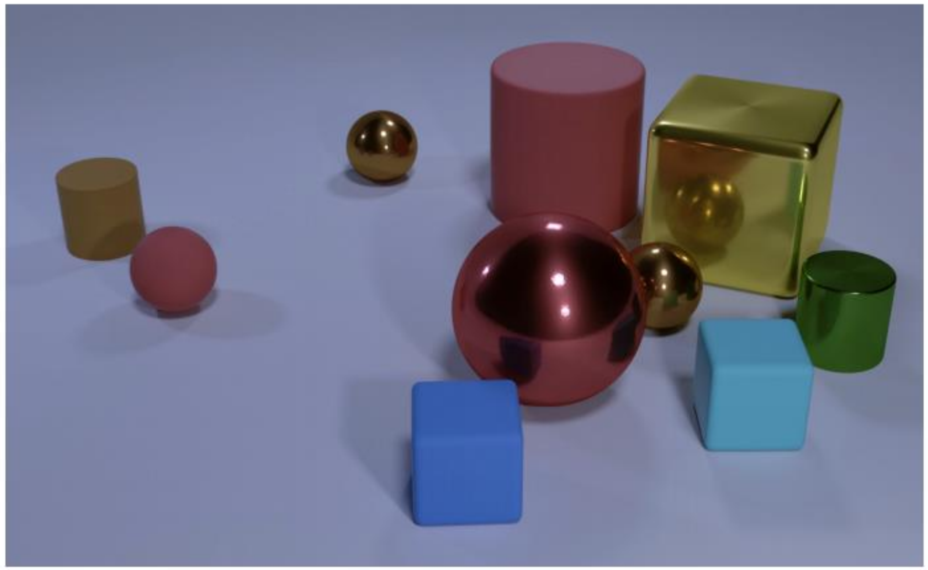
$\texttt{measure} : \mathit{Attention} \to \mathit{Label}$



Andreas et al., Deep Compositional Question Answering with Neural Module Networks, 2016

CLEVR: Dataset for Visual Reasoning

Perfect for a neural module network!



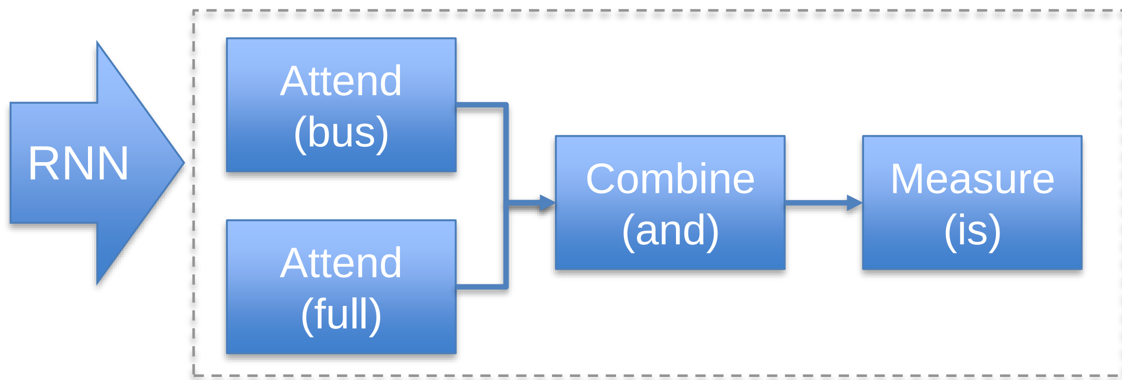
Q: Are there an equal number of large things and metal spheres?
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
Q: How many objects are either small cylinders or metal things?

Johnson et al., CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, CVPR 2017

End-to- End Neural Module Network

Computation layout

Is the bus full of passengers?

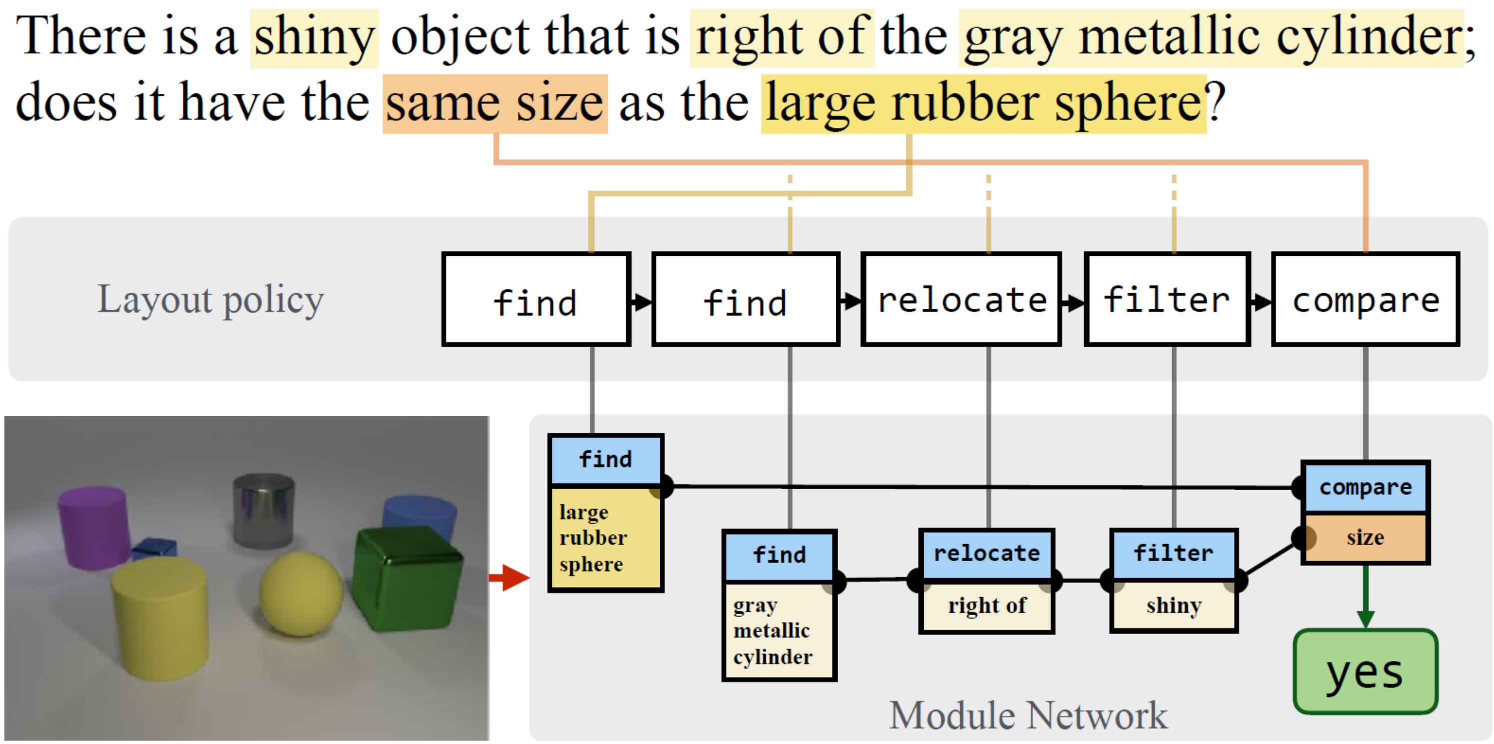


No need to parse the question!

No rule-based creation of the layout!

Hu et al., Learning to Reason: End-to-End Module Networks for Visual Question Answering, 2017

End-to- End Neural Module Network



Hu et al., Learning to Reason: End-to-End Module Networks for Visual Question Answering, 2017