Multimodal Machine Learning

Lecture 7.1:
Alignment and Translation

Louis-Philippe Morency

Administrative Stuff

Student Peer Feedback

Feedback due **Sunday 10/18 at 8pm**(*)

- Student are randomly assigned to 6 videos
  - Learn about your 6 assigned videos: http://atlas.multicomp.cs.cmu.edu:8300/
- Share your feedback using online form: https://forms.gle/Ji5kuLqJwRXC6CTS8
  - Be constructive in your feedback
  - Respect other's ideas. No plagiarism.
- List of video links on Piazza

(*) Friday October 16th is a holiday. Participate in Tartan Community Day

Wild Cards for Project Assignments

Updated late-submission policy:

Each team has two (2) wild cards

- Each wild card allows up to 24 extra hours
- Wild cards can be cumulated, or used separately
- Send a note on Piazza BEFORE using a card
- No partial credits for wild cards
- Can be used for first, midterm or final assignments

(report or presentation deadlines)

Reading and Lecture Assignments

Reading assignments

- 10 reading assignments are planned
- Your final grade = top 8 scores

Lecture highlights assignments

- 20 lecture highlights are planned
- Your final grade = top 16 scores

If you need flexibility, please contact us via Piazza

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 7 10/13 & 10/15	**Alignment and translation** • Neural Module networks • Connectionist temporal classification	**Probabilistic graphical models** • Dynamic Bayesian networks • Coupled and factor HMMs
Week 8 10/20 & 10/22	**Discriminative graphical models** • Conditional random fields • Continuous and fully-connected CRFs	**Neural Generative Models** • Variational auto-encoder • Generative adversarial networks
Week 9 10/27 & 10/29	**Reinforcement learning** • Markov decision process • Q learning and policy gradients	**Multimodal RL** • Deep Q learning • Multimodal applications
Week 10 11/3 & 11/5	**Fusion and co-learning** • Multi-kernel learning and fusion • Few shot learning and co-learning	**New research directions** • Recent approaches in multimodal ML
Week 11 11/10 & 11/12	***Mid-term project assignment*** *(live working sessions instead of lectures)*	

Midterm project assignment
Presentations due Friday 11/13
Reports due Sunday 11/15
Peer feedback due Sunday 11/22

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 12 11/17 & 11/19	**Embodied Language Grounding** • Connecting Language to Action • Guest lecture: Yonatan Bisk	**Multimodal language acquisition** • Learning from multimodal data • Guest lecture: Graham Neubig
Week 13 11/24 & 11/26	***Thanksgiving week*** *(no lectures)*	
Week 14 12/1 & 12/3	**Learning to connect text and images** • Discourse approaches, text & images • Guest lecture: Malihe Alikhani	**Bias and fairness** • Computational ethics • Guest lecture: Yulia Tsvetkov
Week 15 12/8 & 12/10	***Final project assignment*** *(live working sessions instead of lectures)*	

Final project assignment
Presentations due Friday 12/11
Reports due Sunday 12/13

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 7.1: Alignment and Translation

Louis-Philippe Morency

Learning Objectives of Today's Lecture

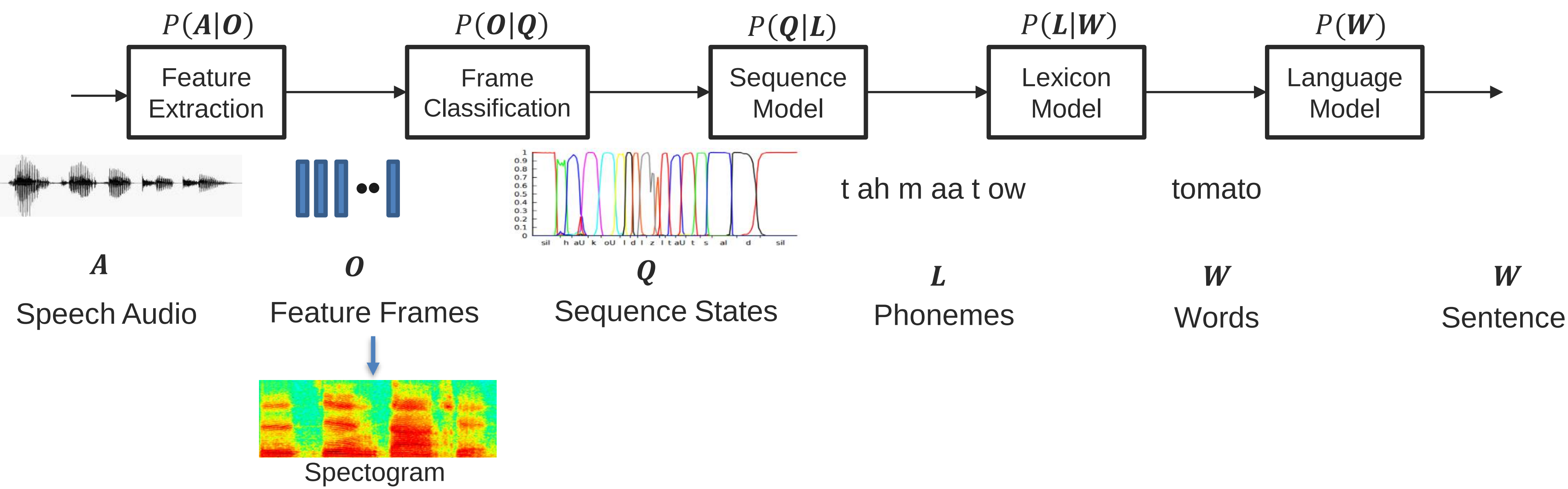
- Multimodal Alignment
  - Alignment for speech recognition
    - Connectionist Temporal Classification (CTC)
  - Multi-view video alignment
    - Temporal Cycle-Consistency
- Multimodal Translation
  - Visual Question Answering
    - Co-attention, Stacked attention
  - Neural module networks
    - Neural-symbolic learning
- Speech-video translation applications
  - Sound of pixels and Speech2Face

Alignment for Speech Recognition

Architecture of Speech Recognition

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

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

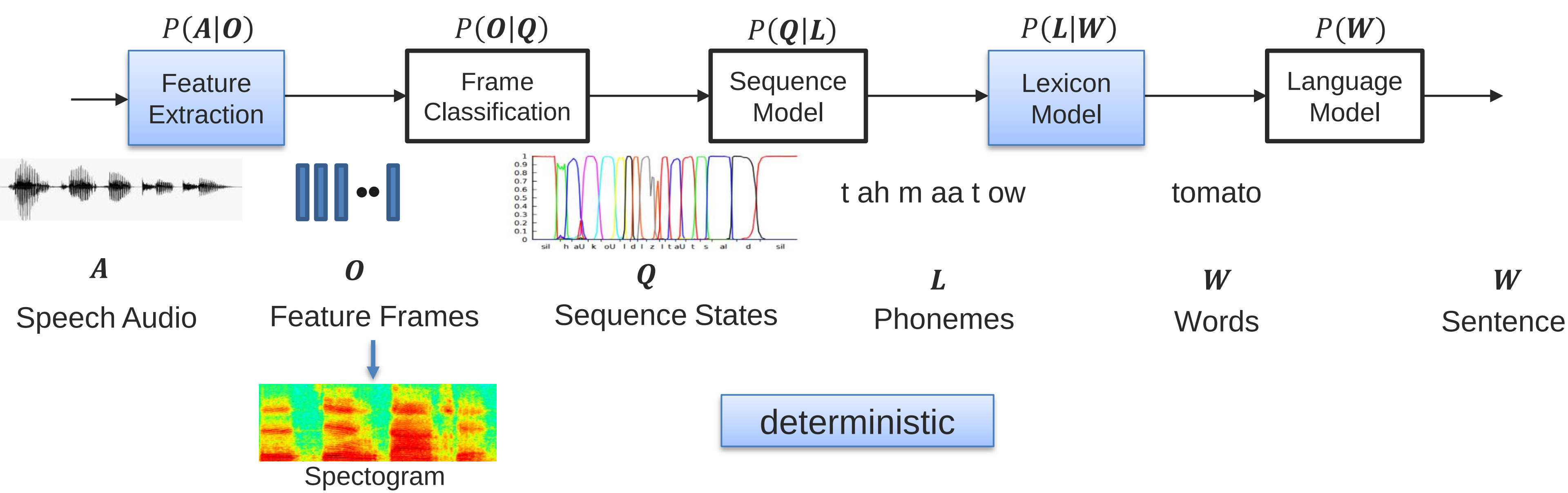


http://slazebni.cs.illinois.edu/spring17/lec26_audio.pdf

Architecture of Speech Recognition

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

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

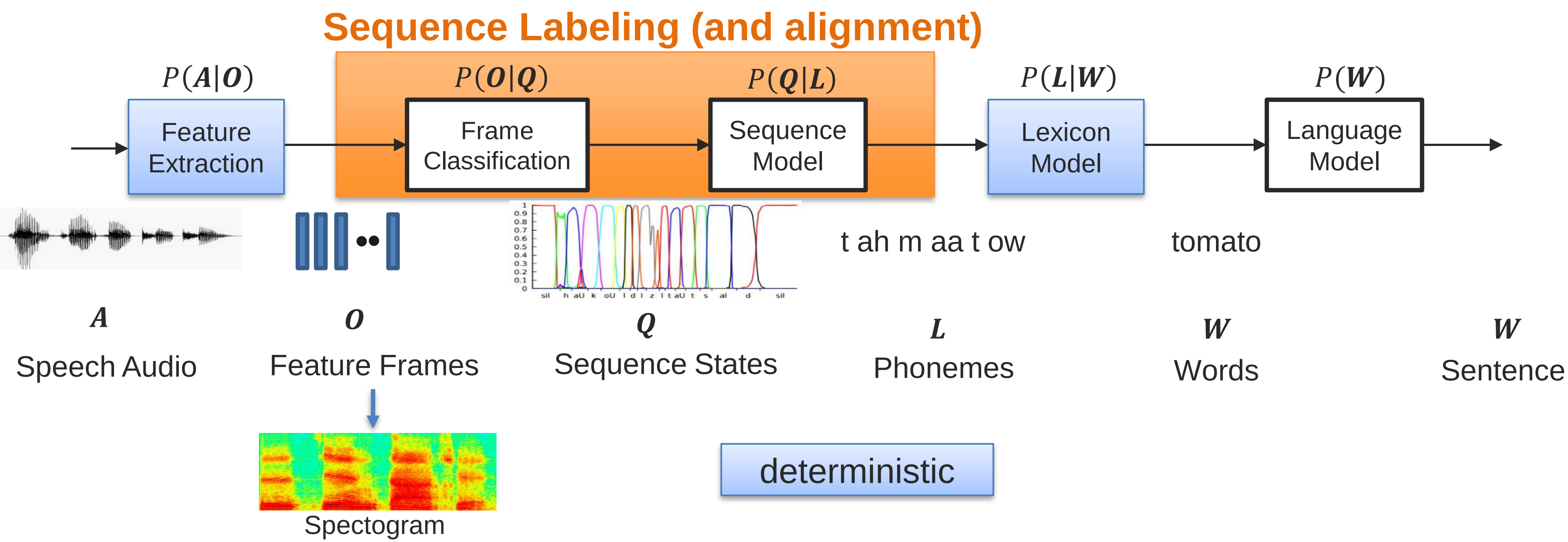


http://slazebni.cs.illinois.edu/spring17/lec26_audio.pdf

Architecture of Speech Recognition

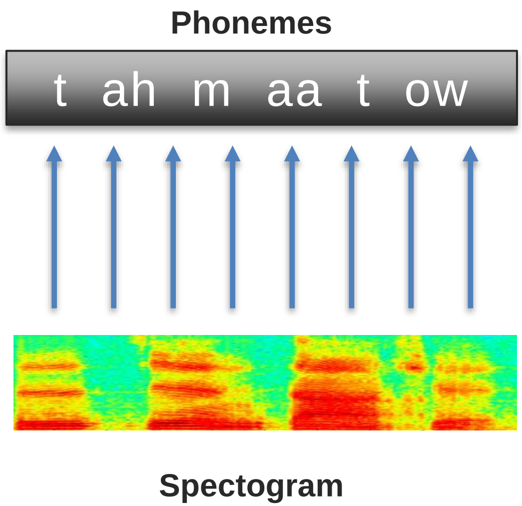
$$\widehat{\boldsymbol{W}} = \underset{\boldsymbol{W}}{\operatorname{argmax}}\, P(\boldsymbol{W}|\boldsymbol{O})$$

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



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

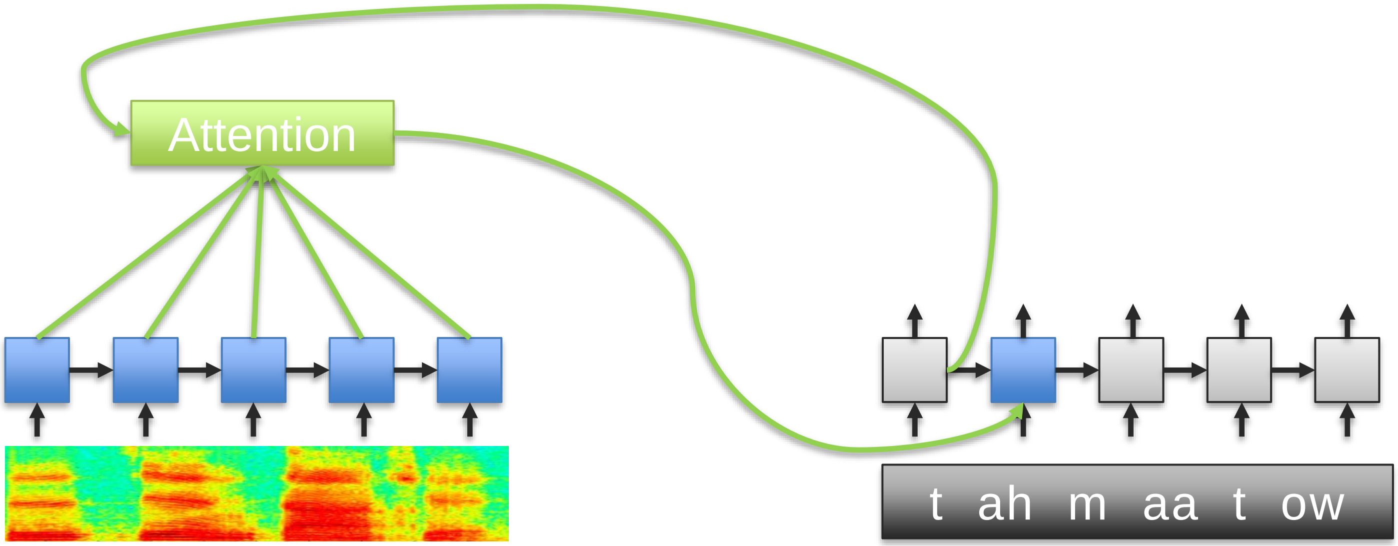
Sequence Labeling (and Alignment)



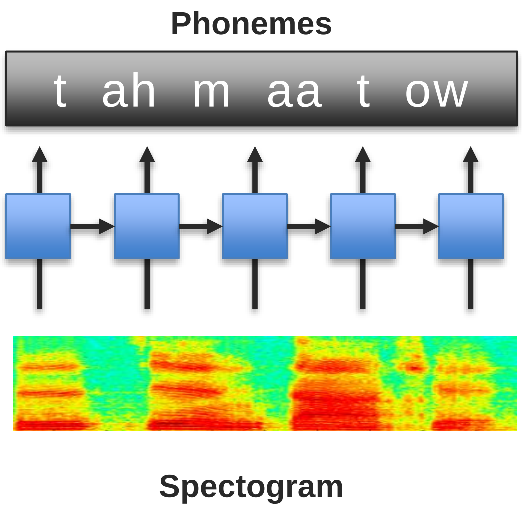
How can we predict the sequence of phoneme labels from the sequence of audio frames?

Option 1: Sequence-to-Sequence (Seq2Seq)

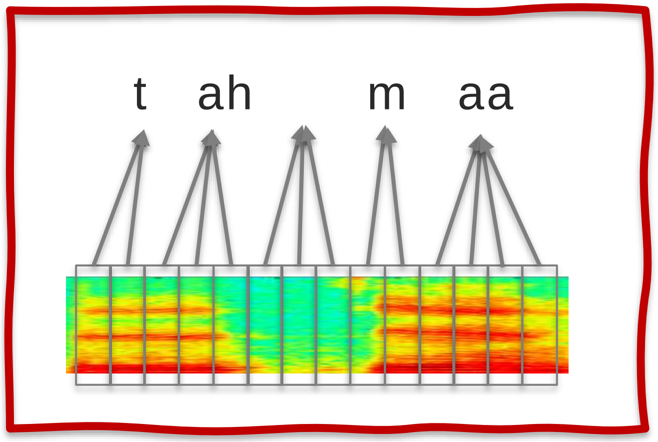
Option 2: Seq2Seq with Attention



Option 3: Sequence Labeling with RNN



Challenge: many-to-1 alignment



What should be the loss function?

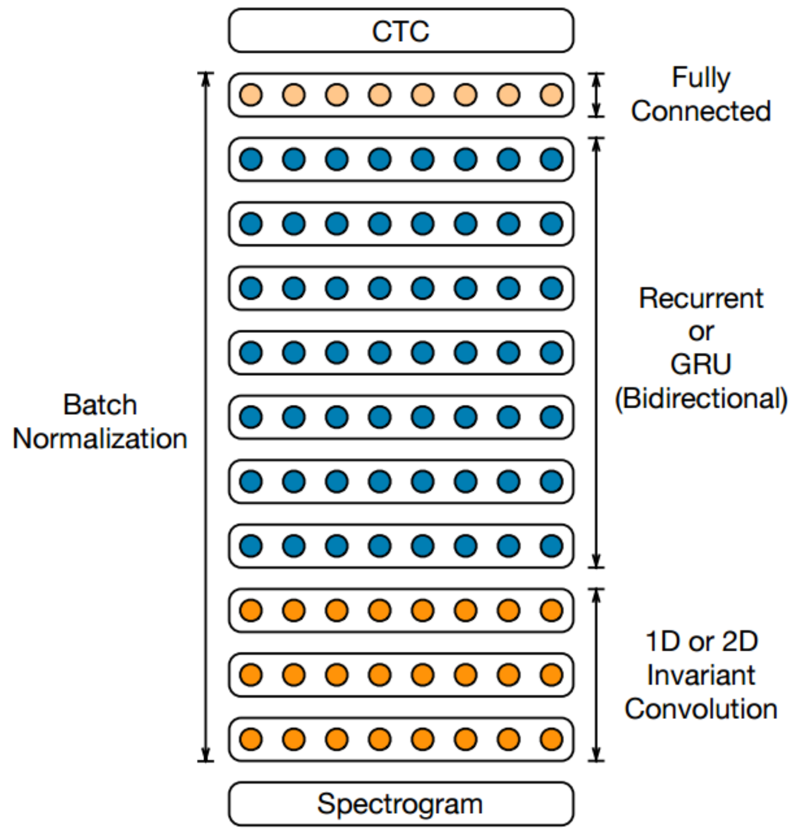
Speech Alignment

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

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

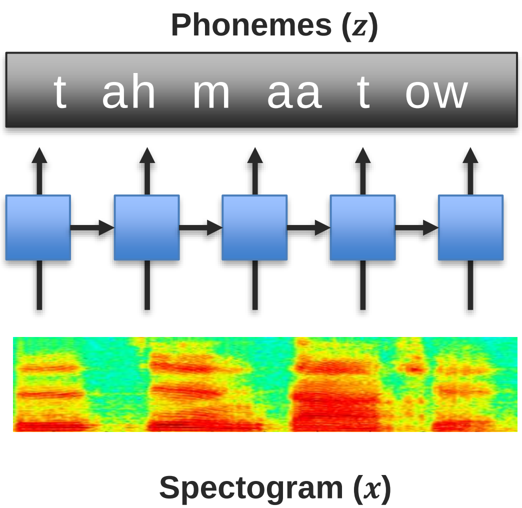
$$\boldsymbol{x} = (x_1, x_2, \dots, x_T)$$

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

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

defined over the space of labels L

Not the same length $U \le T$



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

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 $\boldsymbol{l}$

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

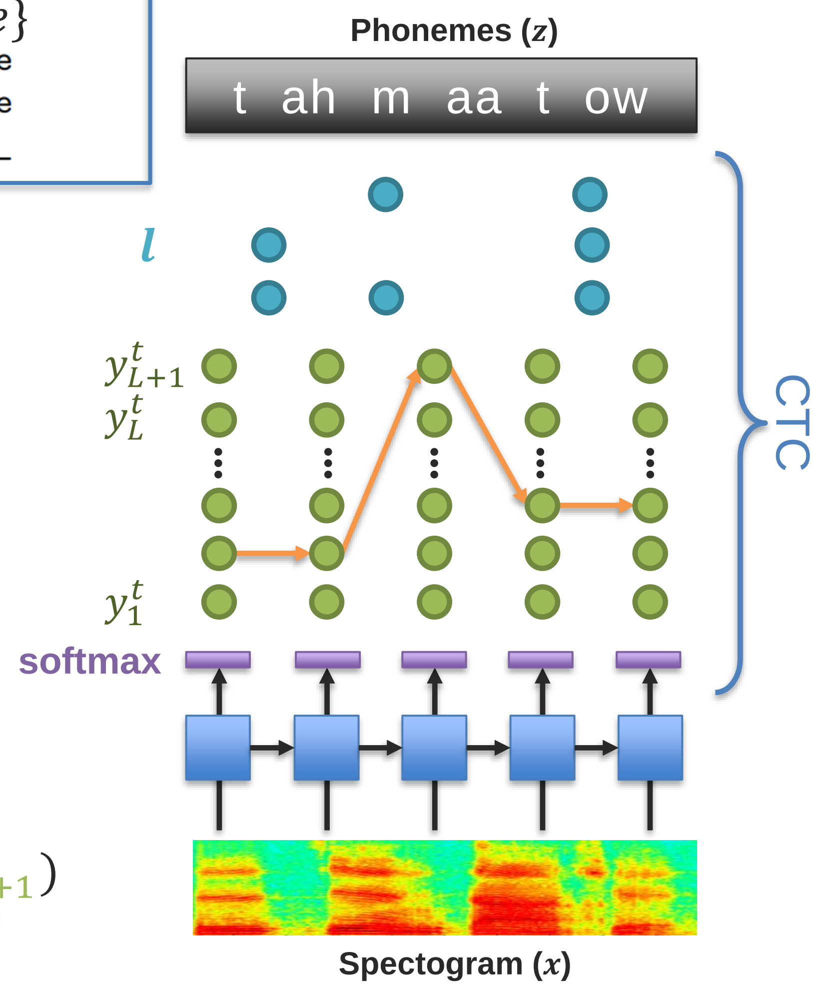
② Path $\boldsymbol{\pi}$ over the activations:

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

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 $\boldsymbol{l}$

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

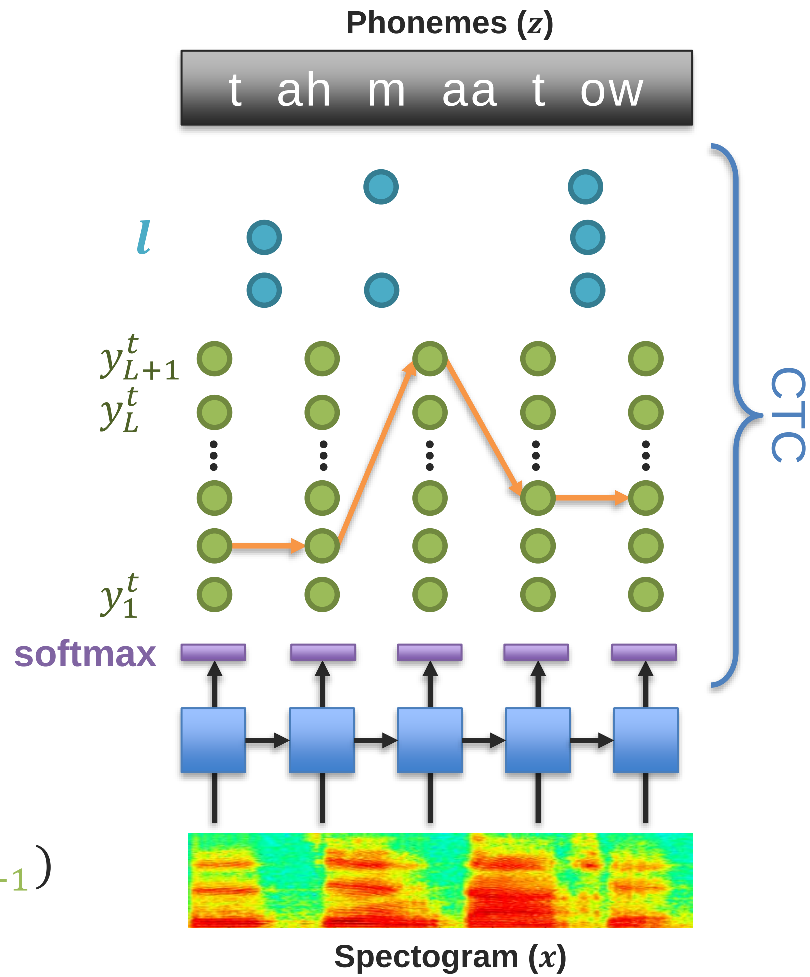
2 Path $\boldsymbol{\pi}$ over the activations:

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

1 Output activations (distribution):

$$\boldsymbol{y} = f_\theta(\boldsymbol{x}), \text{ where } \boldsymbol{y}^t = (y_1^t, y_2^t, \ldots, y_L^t, y_{L+1}^t)$$

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 $\boldsymbol{\pi}$

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

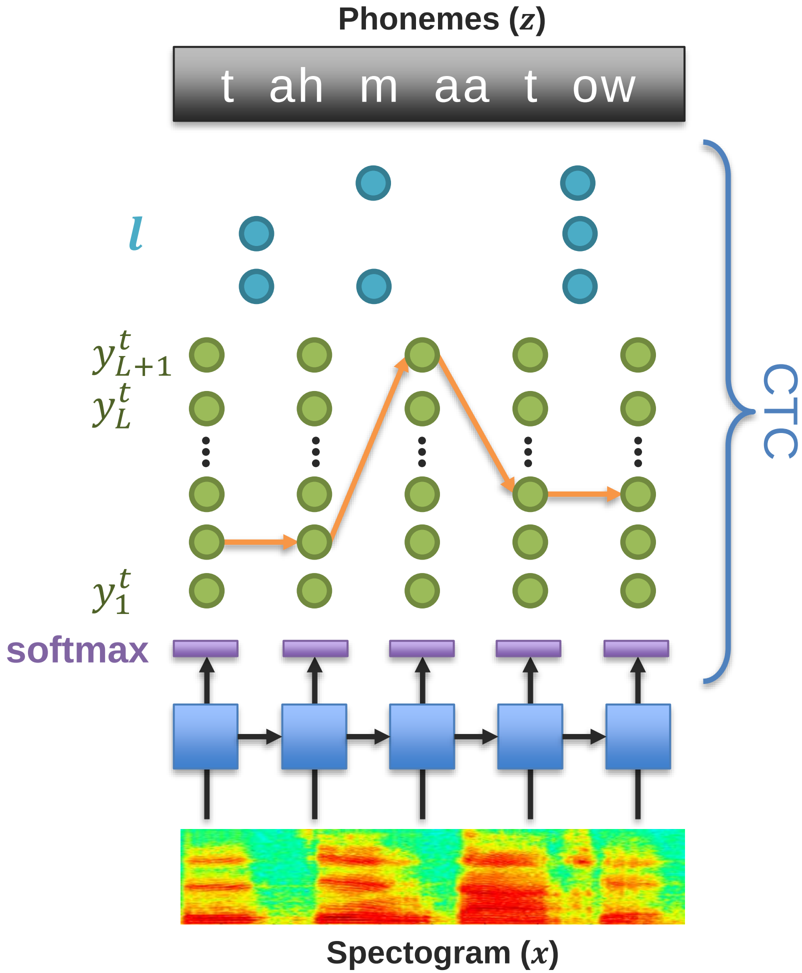
Get most probable labels z^* directly from π^*

Option 2: Solve using dynamic programming

Forward-backward algorithm

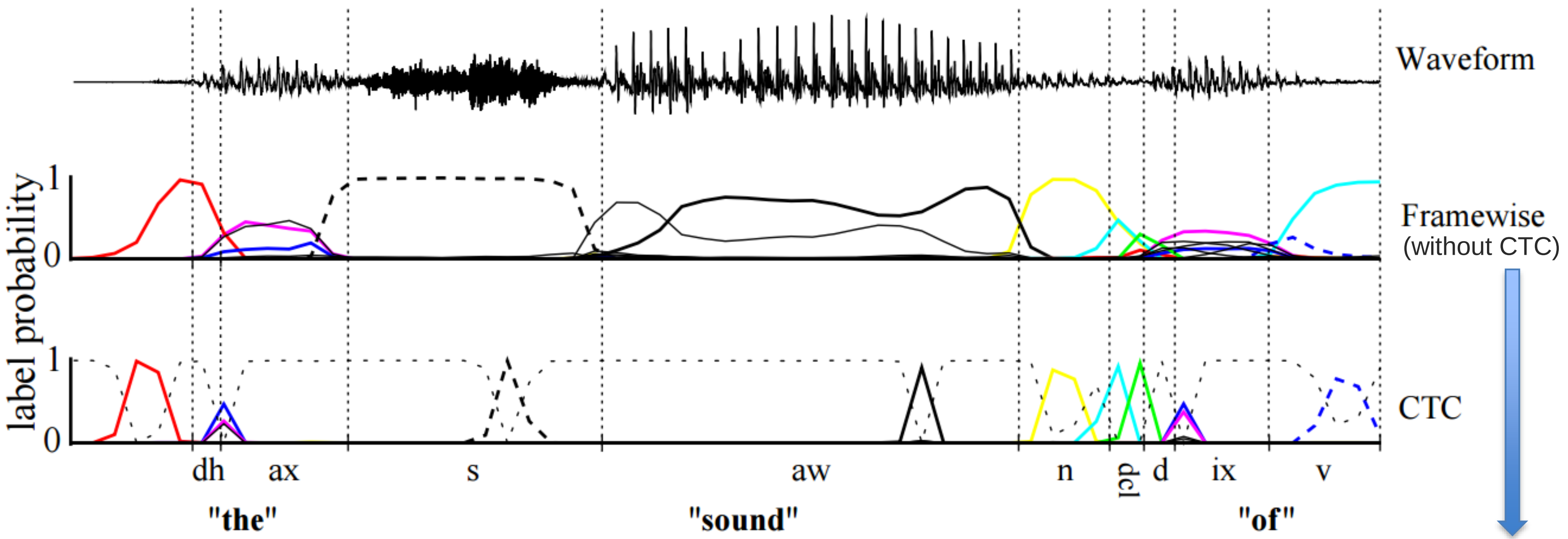
- Forward variables α
- Backward variables β

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



Visualizing CTC Predictions

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



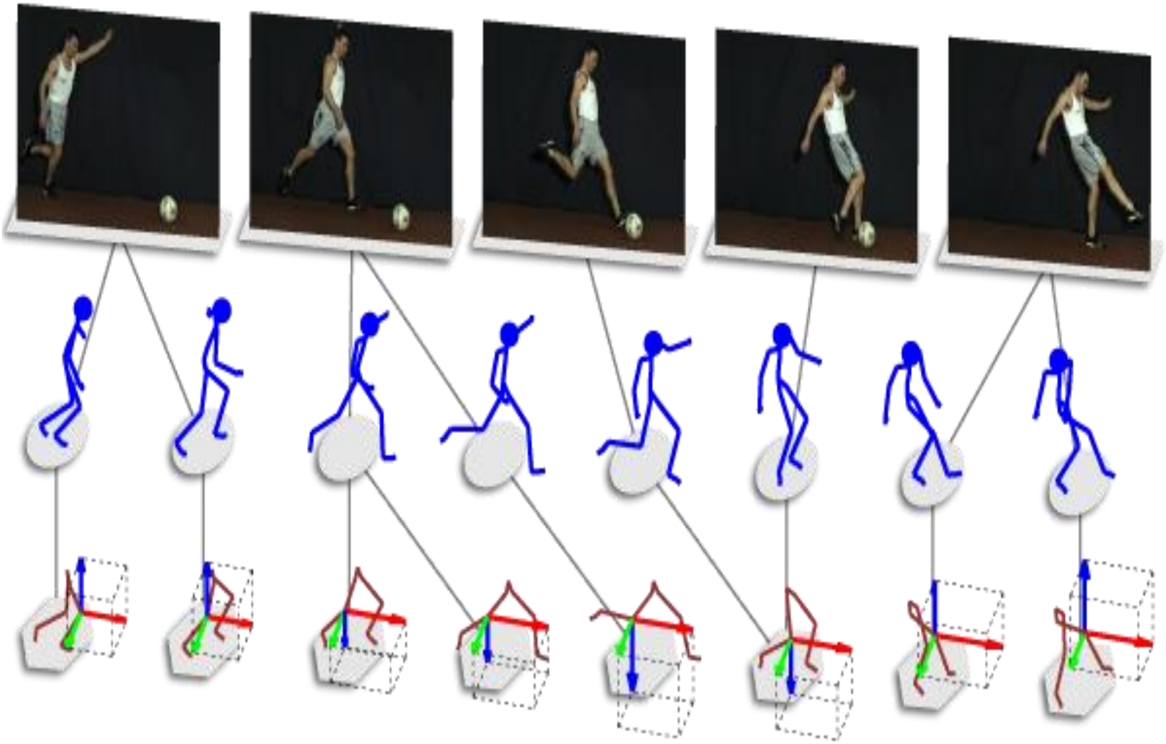
It gets penalized for mistakes around the boundaries

Why are CTC predictions so “peaky”?

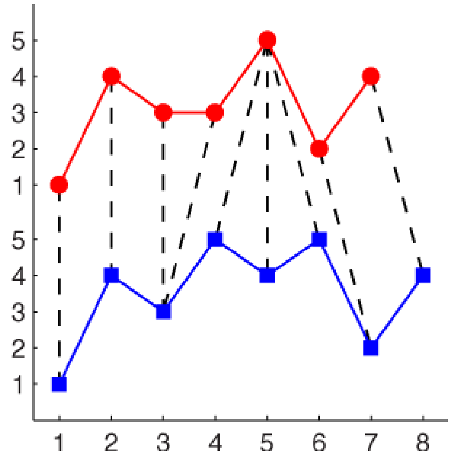
CTC focuses on the phoneme transitions

Multi-View Video Alignment

Temporal sequence alignment

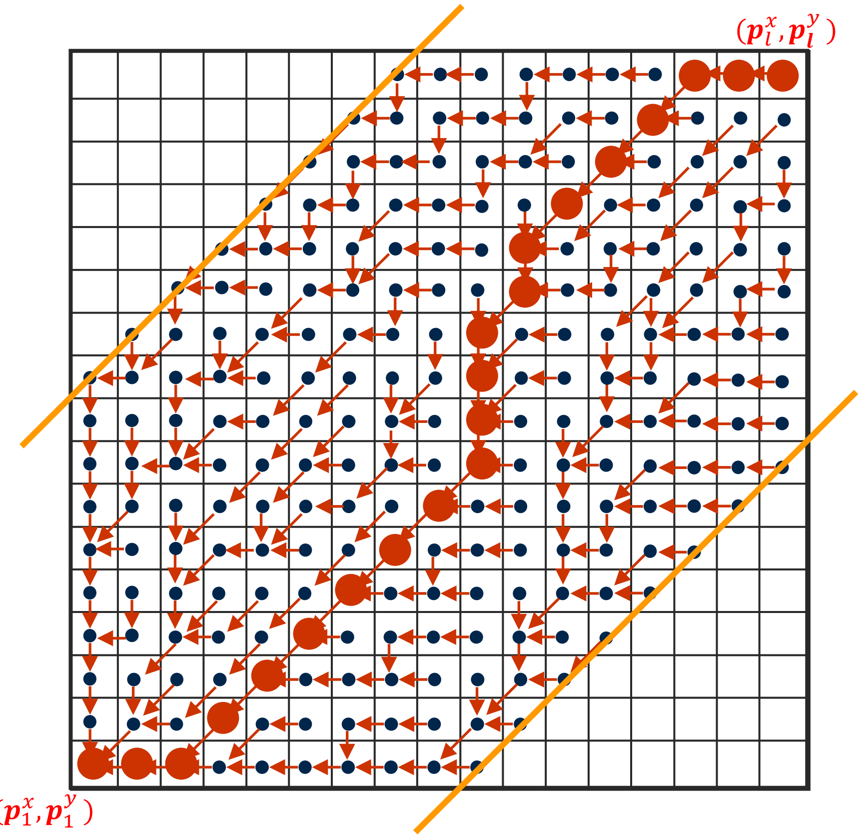


Dynamic Time Warping



$$L(\boldsymbol{p}_t^x, \boldsymbol{p}_t^y) = \sum_{t=1}^{l} \left\| \boldsymbol{x}_{\boldsymbol{p}_t^x} - \boldsymbol{y}_{\boldsymbol{p}_t^y} \right\|_2^2$$

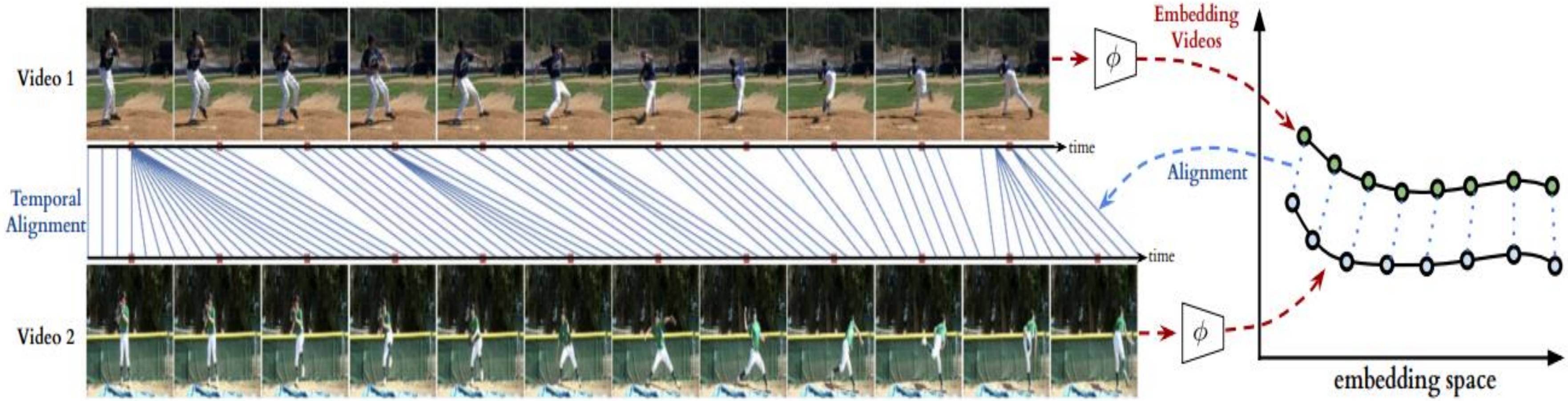
Solved with dynamic programming…



A differentiable version of DTW also exists…
This is one of the reading assignment this week!

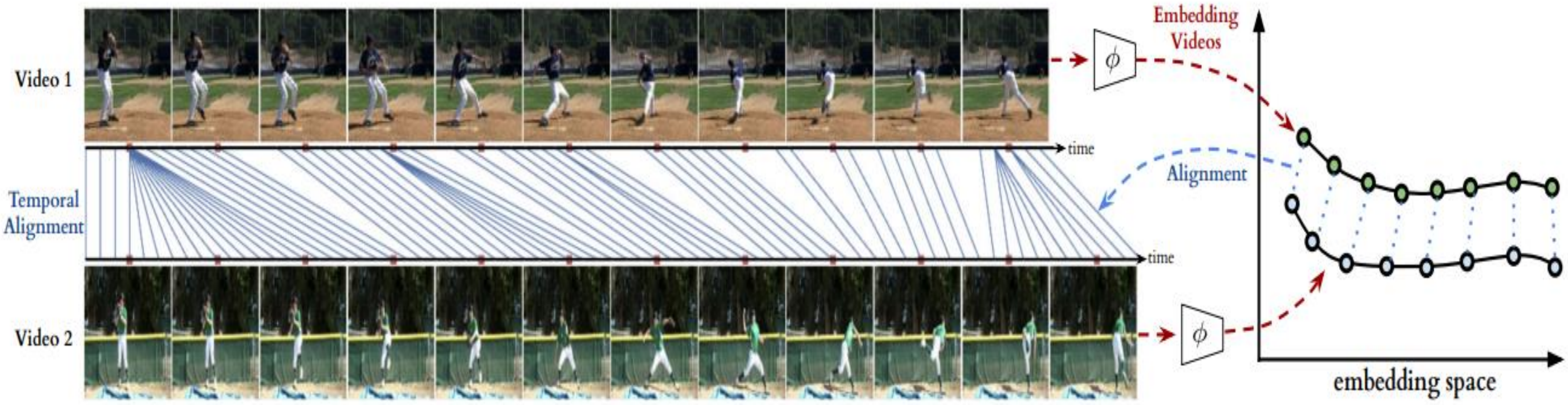
Temporal Alignment using Neural Representations

Premise: we have paired video sequences that can be be temporally aligned



How can we define a loss function to enforce the alignment between sequences?

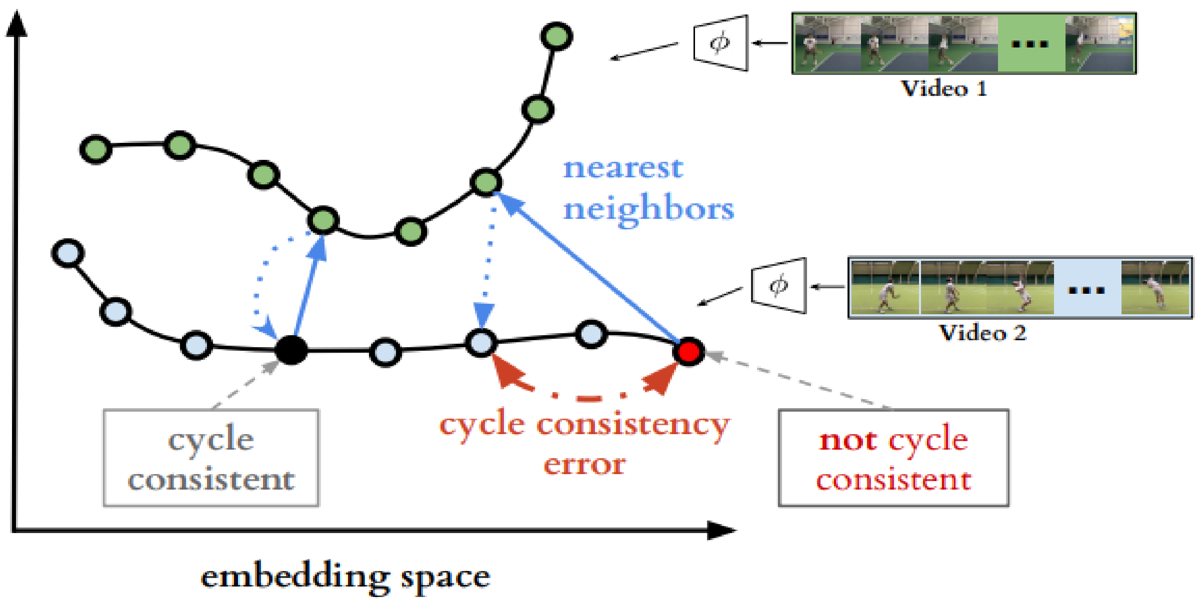
Temporal Cycle-Consistency Learning



Self-supervised approach to learn an embedding space where two similar video sequences can be aligned temporally

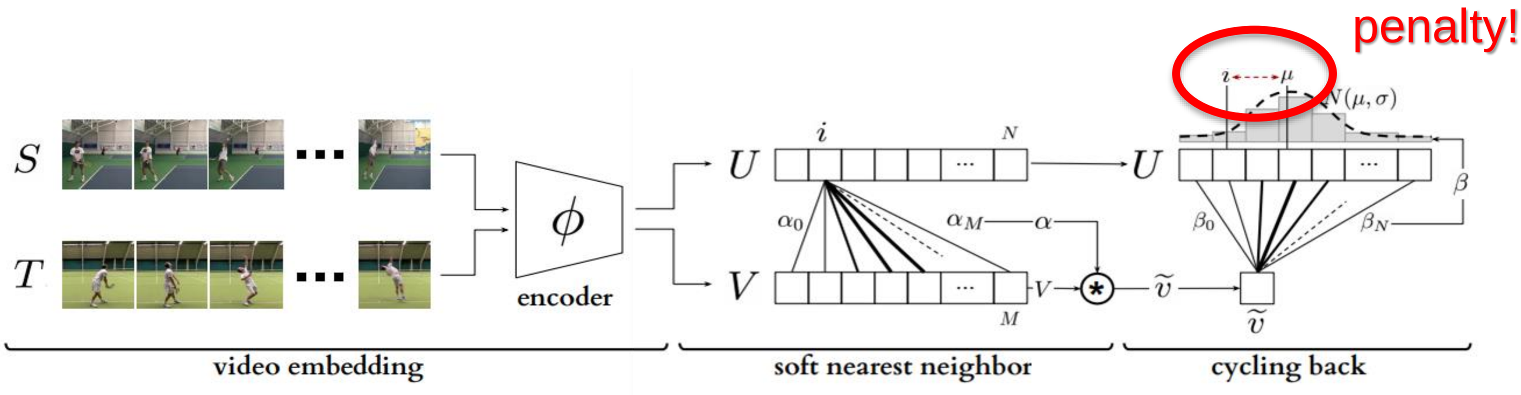
Temporal Cycle-Consistency Learning

Representation learning by enforcing **Cycle consistency**



Main idea: My closest neighbor should also be your closest neighbor

Temporal Cycle-Consistency Learning



Compute “soft” / “weighted” nearest neighbour:

distances: $$\alpha_j = \frac{e^{-||u_i - v_j||^2}}{\sum_k^M e^{-||u_i - v_k||^2}}$$ Soft nearest neighbor: $$\widetilde{v} = \sum_j^M \alpha_j v_j,$$

Find the nearest neighbor the other way and then penalize the distance:

$$\beta_k = \frac{e^{-||\widetilde{v} - u_k||^2}}{\sum_j^N e^{-||\widetilde{v} - u_j||^2}}$$

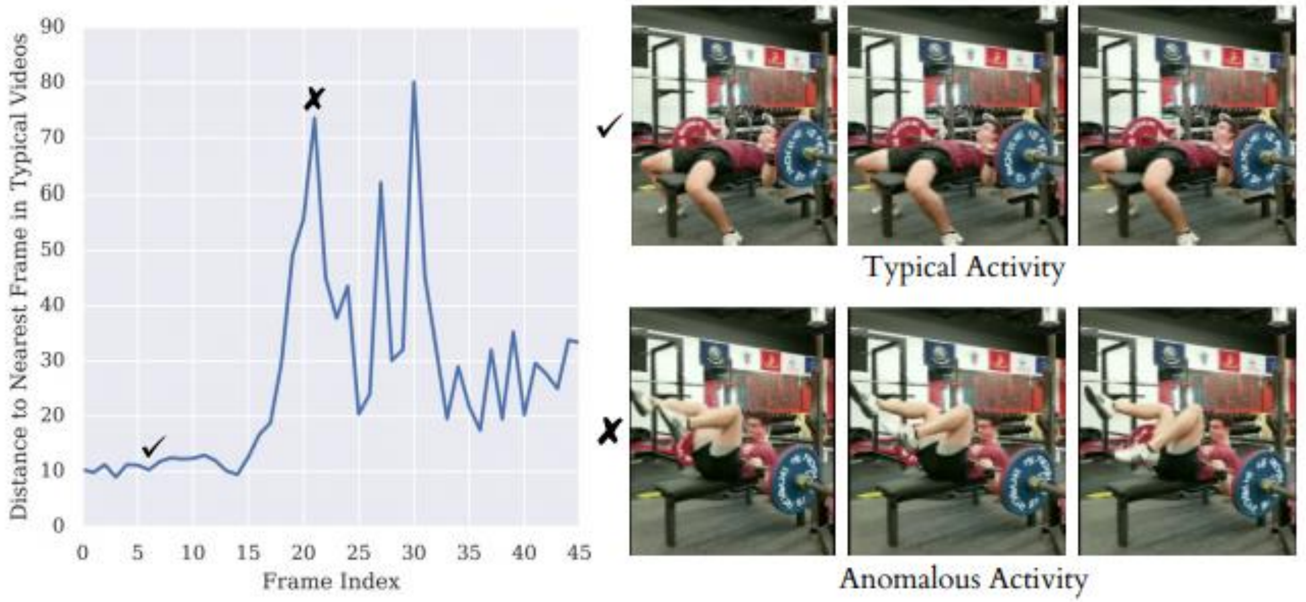
$$L_{cbr} = \frac{|i-\mu|^2}{\sigma^2} + \lambda \log(\sigma)$$

Temporal Cycle-Consistency Learning

Nearest Neighbour Retrieval

Temporal Cycle-Consistency Learning

Anomaly Detection



How could you extend this idea to multimodal?

Multimodal Translation
Visual Question Answering (VQA)

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

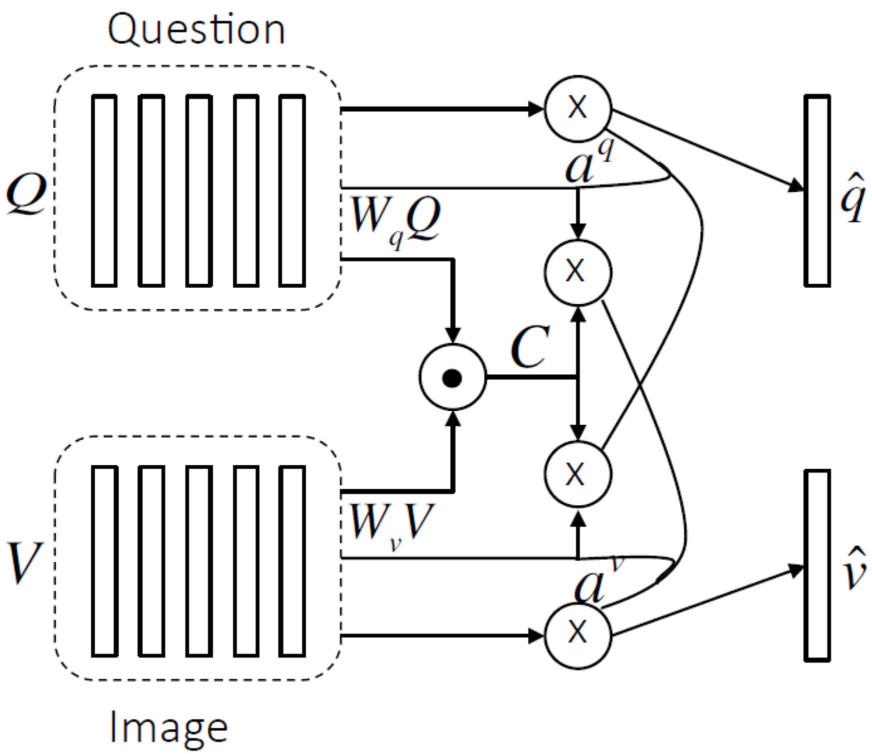
Lu et al., Hierarchical Question-Image Co-Attention for Visual Question Answering, NIPS 2016

Co-attention

Question

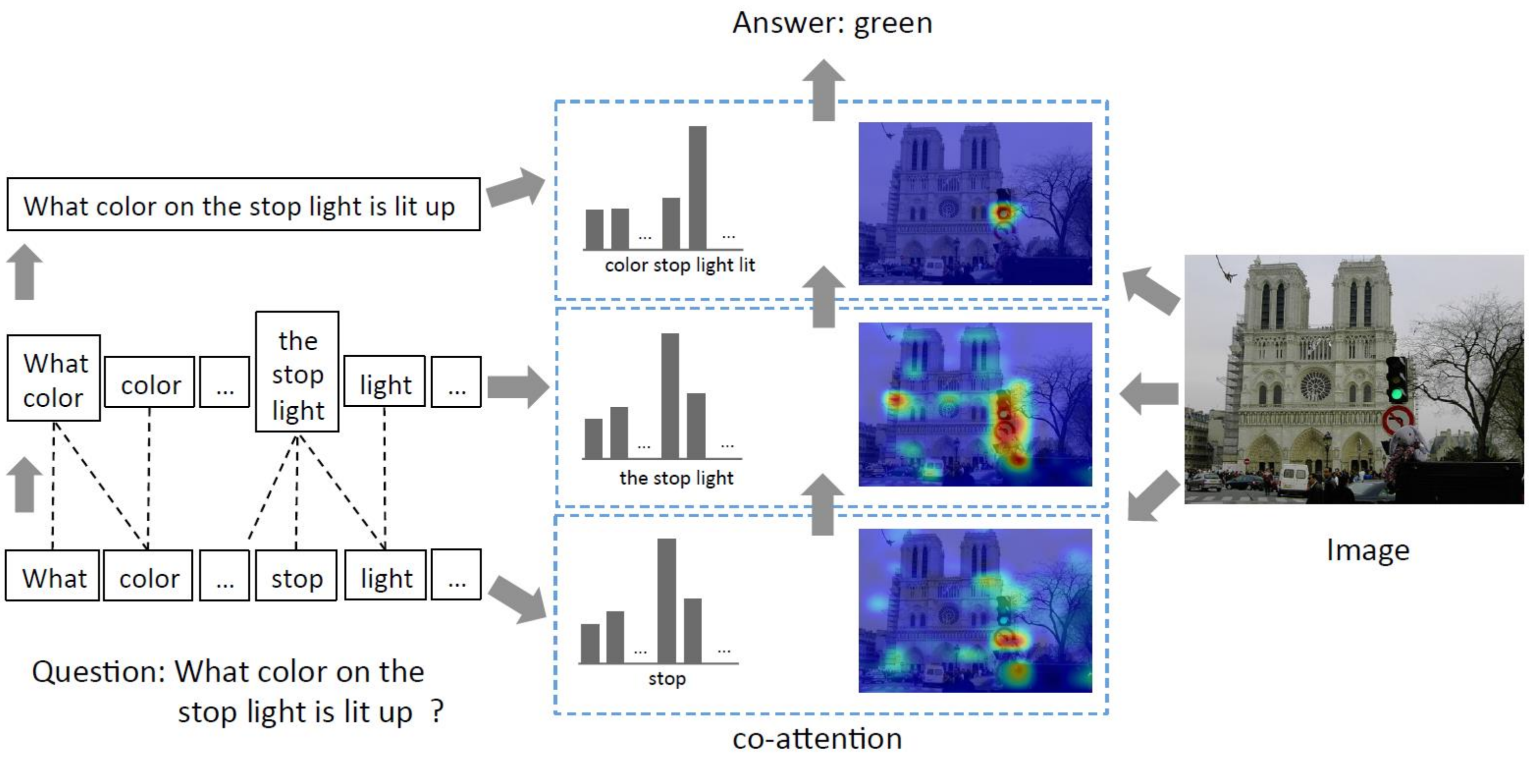
Is the skateboard airborne?

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

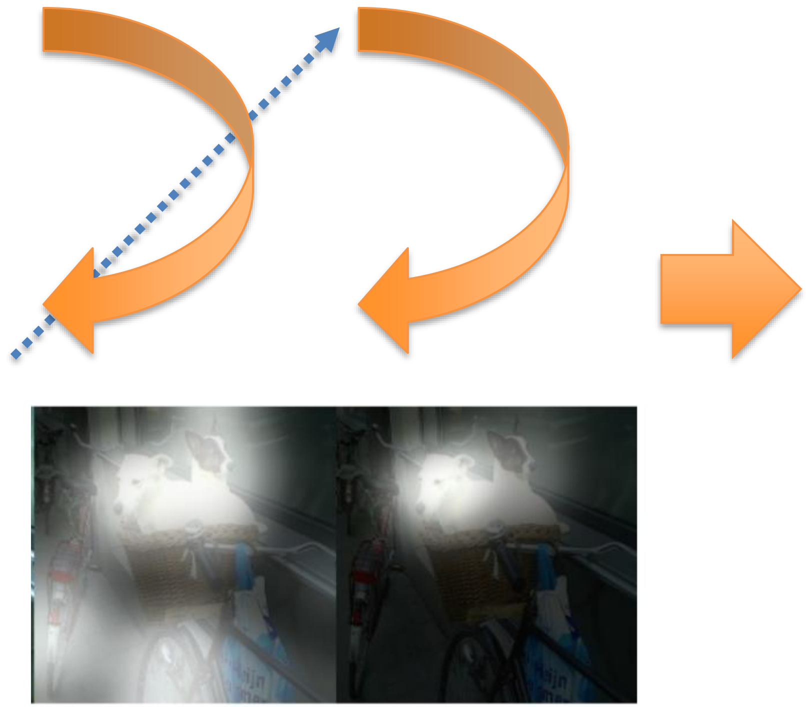
Stacked Attentions

Question

What are sitting in the basket on a bicycle?

Image

Attention 1 Attention 2



Attention 1 Attention 2

Answer

dogs

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

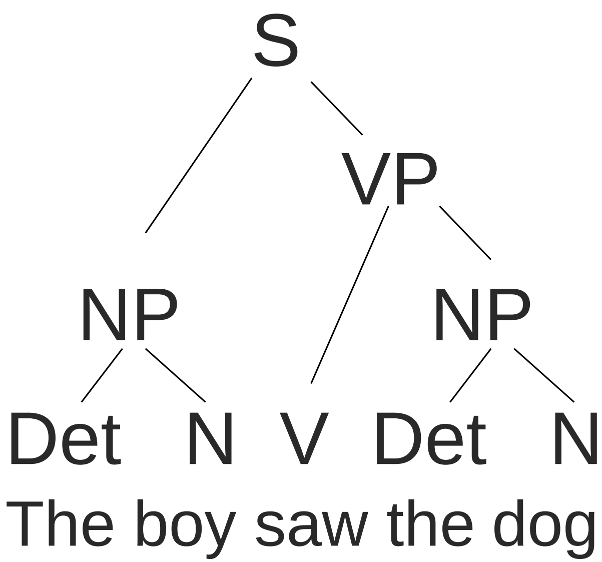
Other Attention-based Models for VQA

- 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

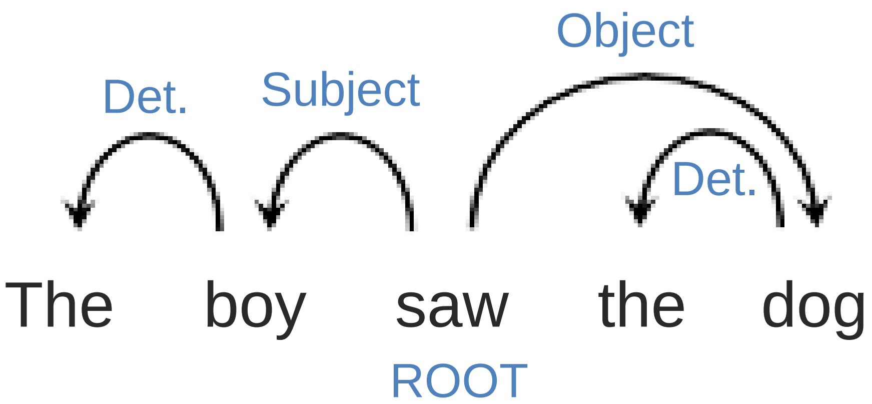
But how to take advantage of language syntax?

VQA: Neural Module Networks

Syntax and Language Structure

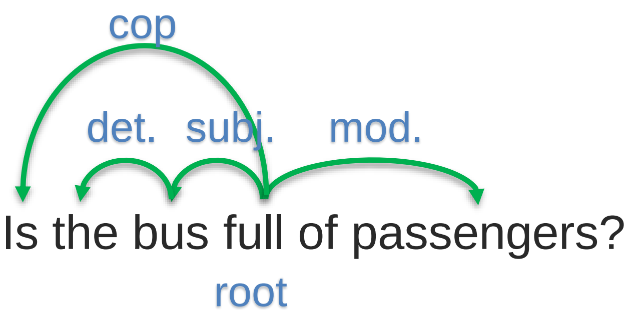


Constituency Parsing

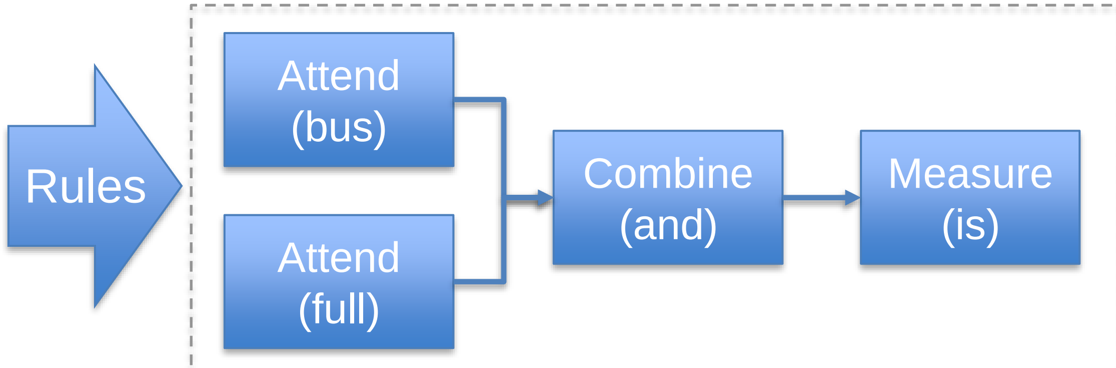


Dependency Parsing

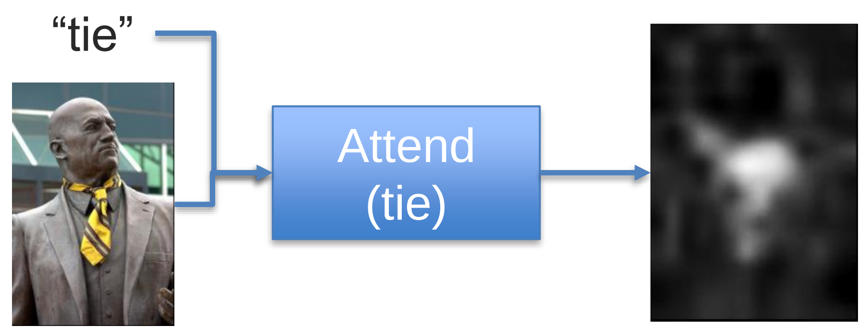
Neural Module Network



Computation layout



Each module work on the attention map(s):

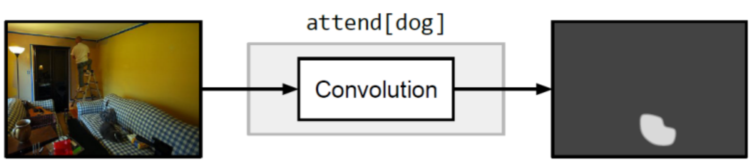


Andreas et al., Deep Compositional Question Answering with Neural Module Networks, 2016

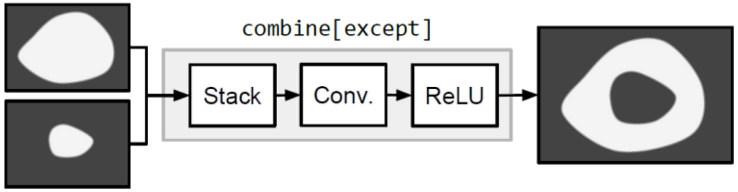
Predefined Set of Modules

1) Analyze the image:

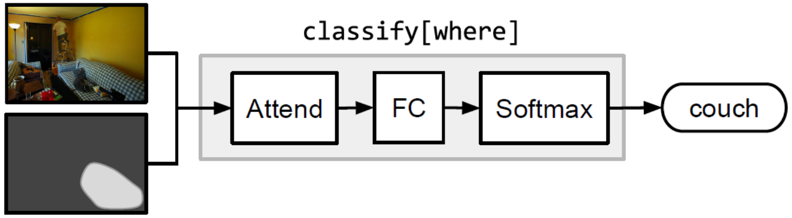
attend : $Image \rightarrow Attention$



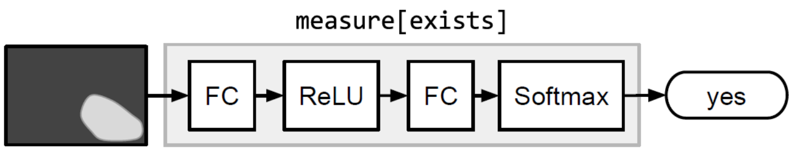
combine : $Attention \times Attention \rightarrow Attention$



2) Make a prediction



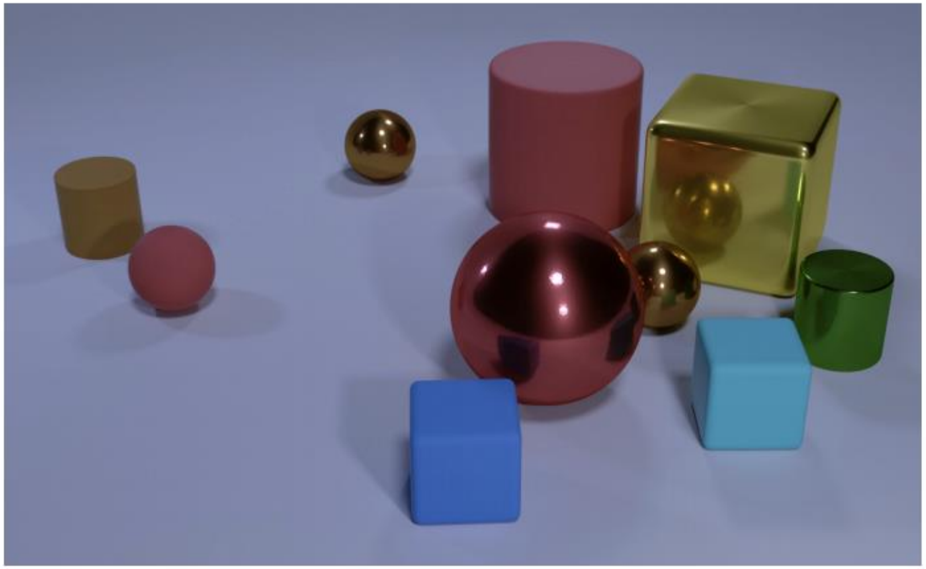
measure : $Attention \rightarrow Label$



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

CLEVR: Dataset for Visual Reasoning

Perfect for a neural module network!



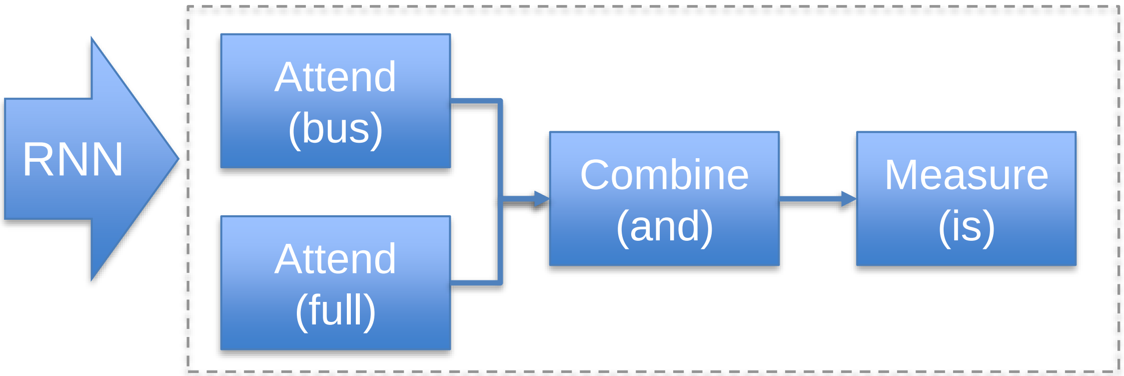
Q: Are there an equal number of large things and metal spheres? **Q:** What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere? **Q:** How many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

Is the bus full of passengers?

Computation layout

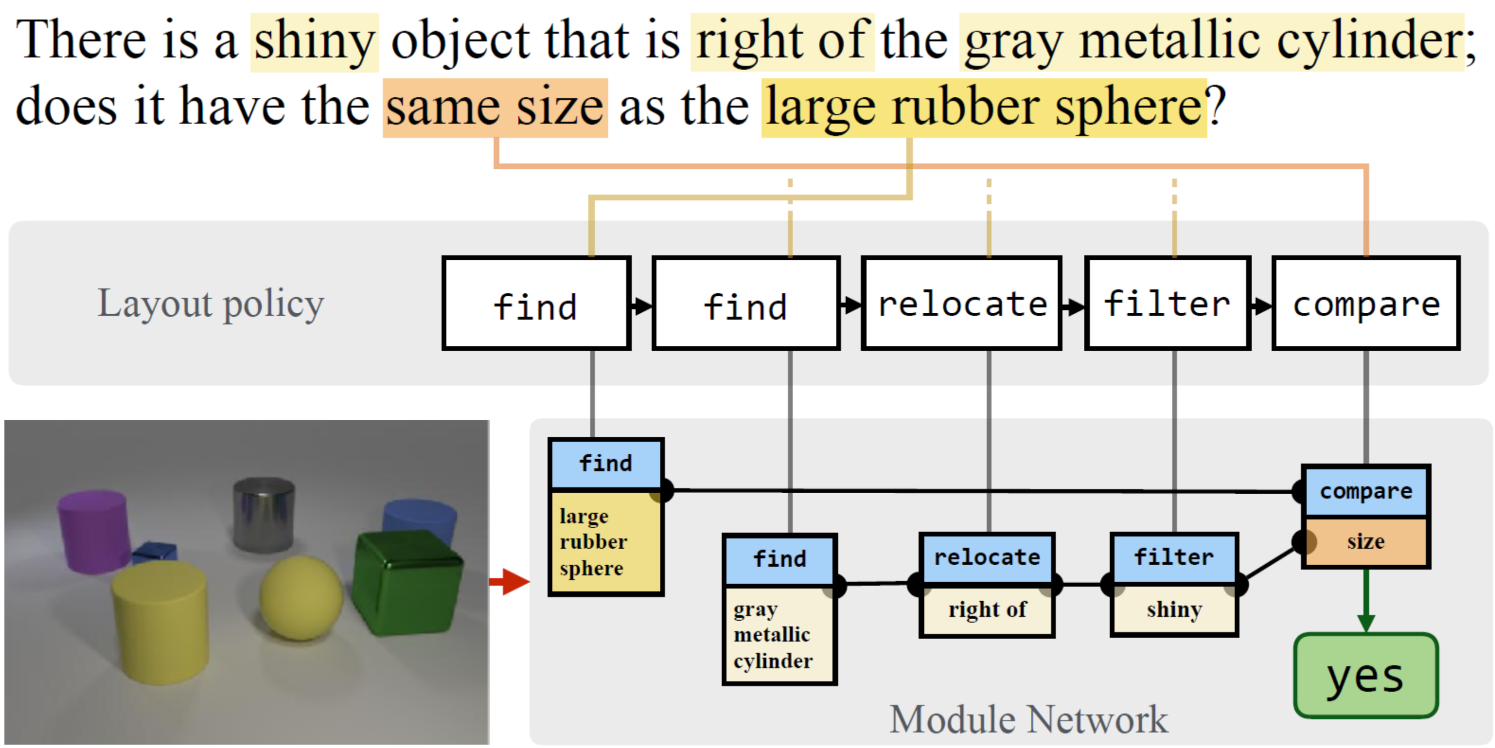


No need to parse the question!

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



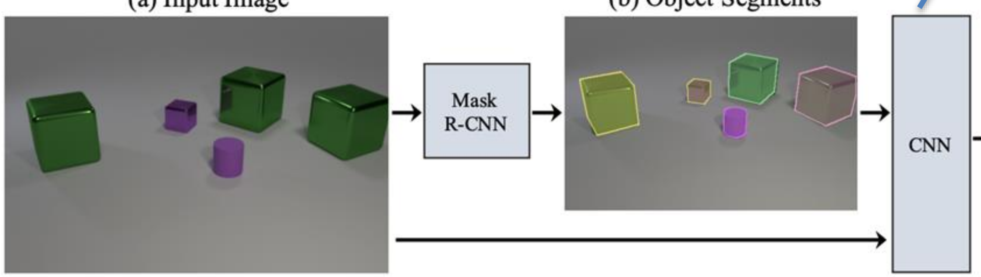
Hu et al., Learning to Reason: End-to-End Module Networks for Visual Question Answering, 2017

VQA: Neural-Symbolic Networks

Neural-symbolic VQA

1) Image de-rendering

Previously trained in a supervised way



(c) Structural Scene Representation

ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

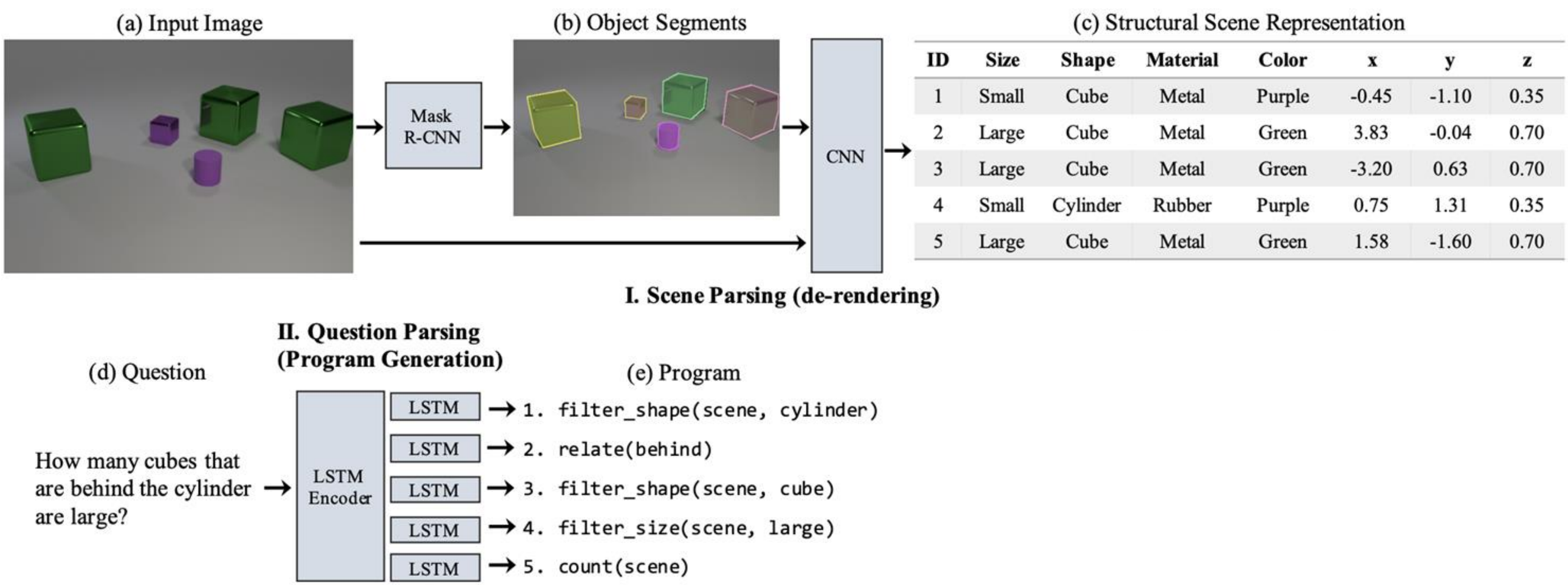
I. Scene Parsing (de-rendering)

Kexin Yi, et al. "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding." Neurips 2018

Neural-symbolic VQA

2) Parsing questions into programs

Similar to neural module networsk



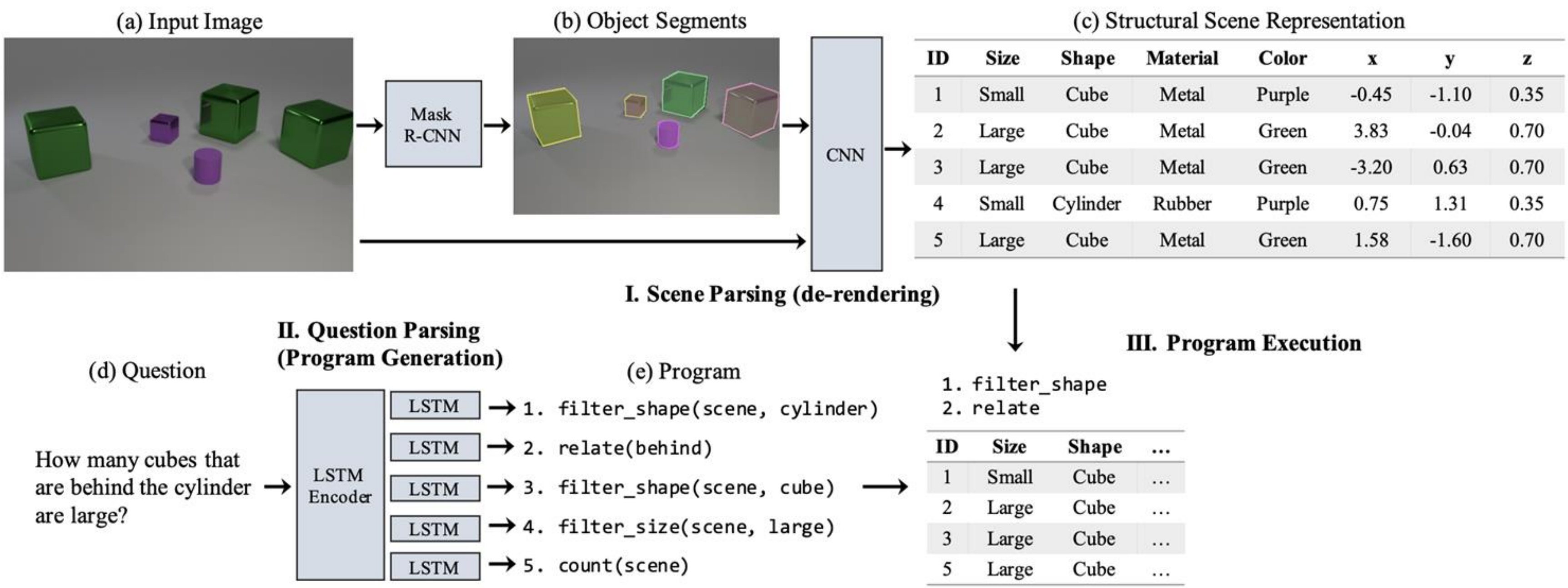
ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

Kexin Yi, et al. “Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding.” Neurips 2018

Neural-symbolic VQA

3) Program execution

Execution of the program is somewhat easier given the “symbolic” representation of the image

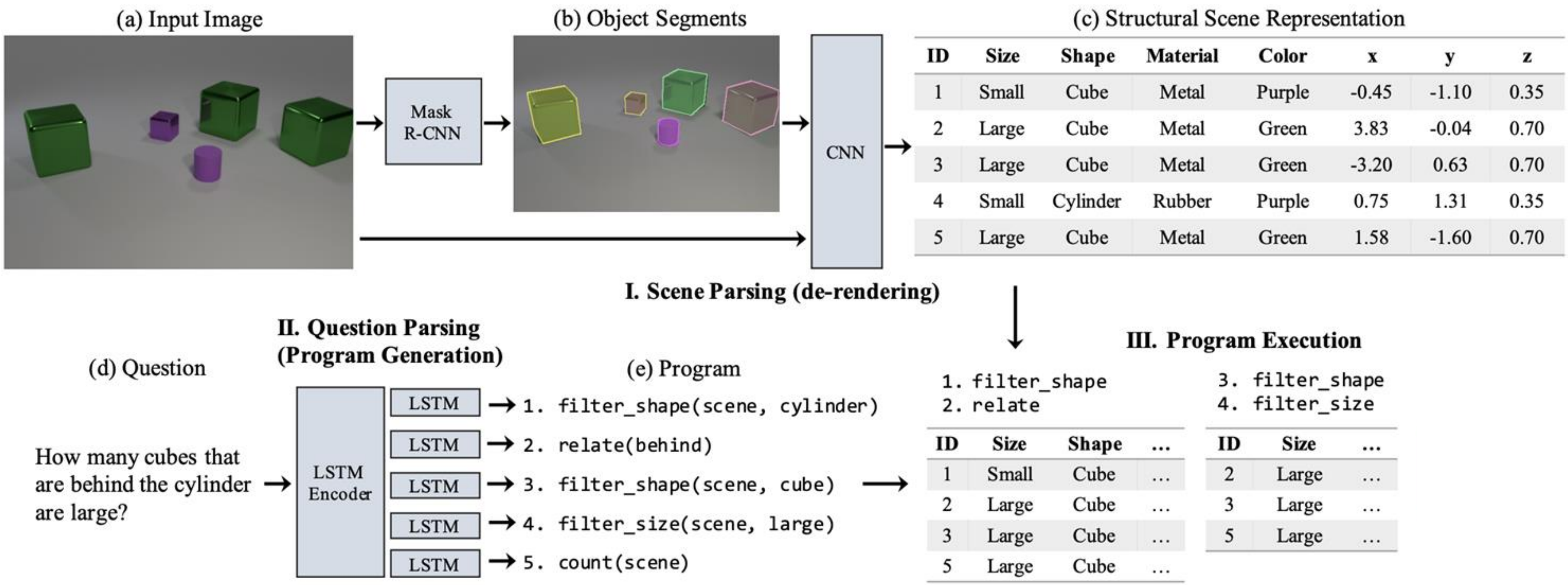


Kexin Yi, et al. “Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding.” Neurips 2018

Neural-symbolic VQA

3) Program execution

Execution of the program is somewhat easier given the "symbolic" representation of the image



ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

ID	Size	Shape	...
1	Small	Cube	...
2	Large	Cube	...
3	Large	Cube	...
5	Large	Cube	...

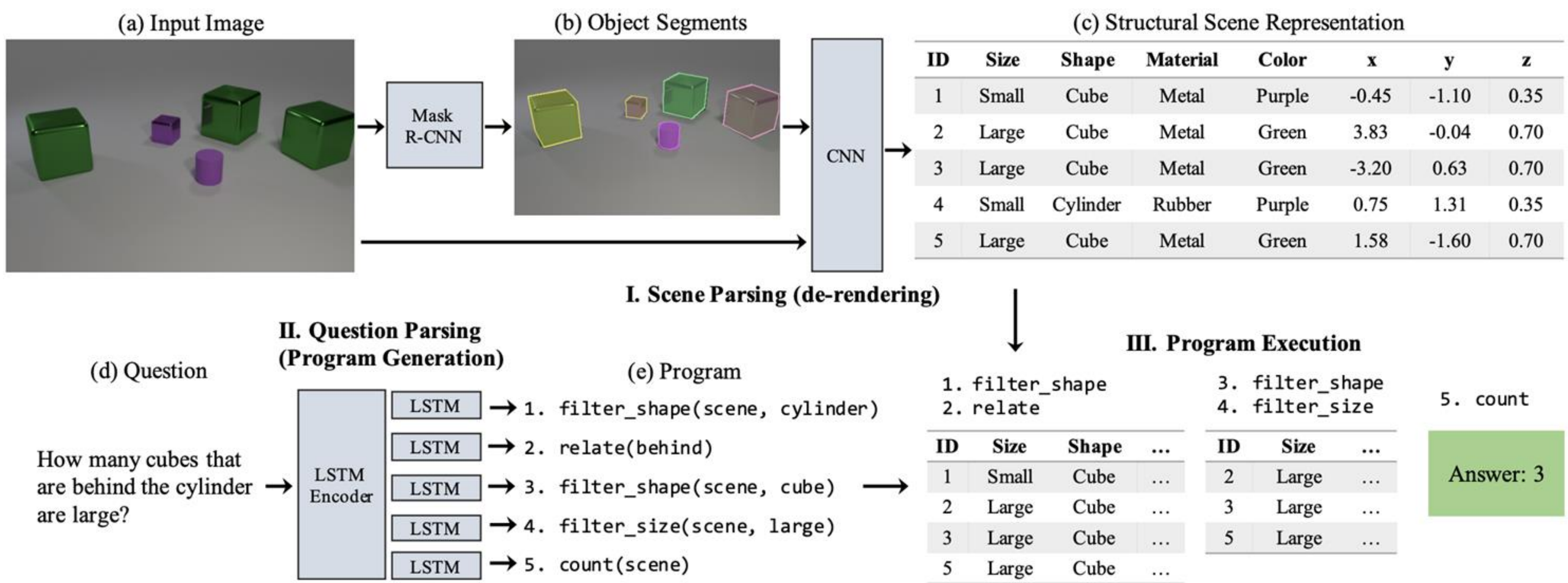
ID	Size	...
2	Large	...
3	Large	...
5	Large	...

Kexin Yi, et al. "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding." Neurips 2018

Neural-symbolic VQA

3) Program execution

Execution of the program is somewhat easier given the "symbolic" representation of the image



ID	Size	Shape	Material	Color	x	y	z
1	Small	Cube	Metal	Purple	-0.45	-1.10	0.35
2	Large	Cube	Metal	Green	3.83	-0.04	0.70
3	Large	Cube	Metal	Green	-3.20	0.63	0.70
4	Small	Cylinder	Rubber	Purple	0.75	1.31	0.35
5	Large	Cube	Metal	Green	1.58	-1.60	0.70

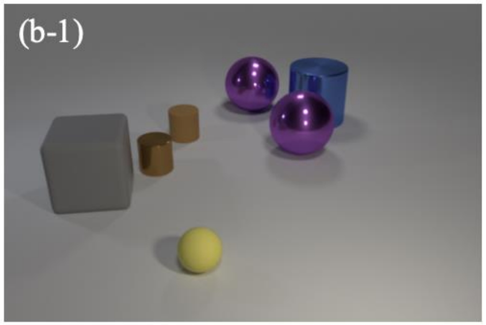
ID	Size	Shape	...
1	Small	Cube	...
2	Large	Cube	...
3	Large	Cube	...
5	Large	Cube	...

ID	Size	...
2	Large	...
3	Large	...
5	Large	...

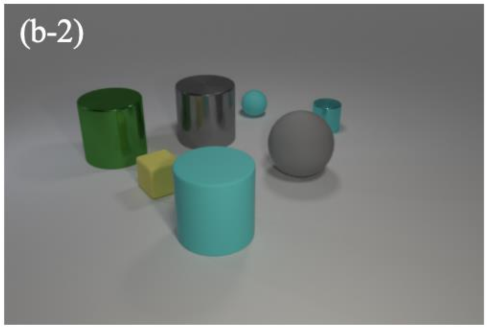
Kexin Yi, et al. "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding." Neurips 2018

Neural-symbolic VQA

(b) 1K Programs



Q: What number of cylinders are gray objects or tiny brown matte objects?



Q: Are there more yellow matte things that are right of the gray ball than cyan metallic objects?

Ours	IEP	Ours	IEP
scene	filter_small	scene	filter_small
filter_small	filter_brown	filter_cyan	filter_cyan
filter_brown	filter_large	filter_metal	union
filter_rubber	filter_cyan	Count	filter_brown
scene	...(25 modules)	...(4 modules)	...(25 modules)
filter_gray	filter_metal	scene	filter_small
union	union	filter_yellow	filter_yellow
filter_cylinder	filter_cylinder	filter_rubber	filter_rubber
count	count	count	count
		greater_than	greater_than
A: 1	**A:** 2	**A:** no	**A:** no

Neural-symbolic programs give more accurate answers

(shown in blue)

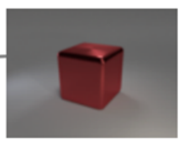
Kexin Yi, et al. "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding." Neurips 2018

The Neuro-symbolic Concept Learner

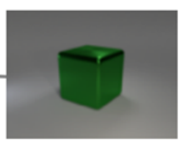
Extension from Neural-symbolic VQA:

Learns visual concepts, words, and semantic parsing of sentences without explicit supervision on any of them, but just by looking at **images and reading paired questions and answers**

I. Learning basic, object-based concepts.

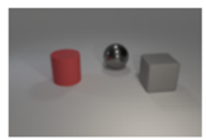


Q: What's the color of the object?
A: Red.
Q: Is there any cube?
A: Yes.



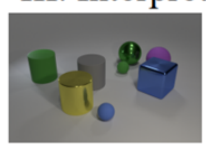
Q: What's the color of the object?
A: Green.
Q: Is there any cube?
A: Yes.

II. Learning relational concepts based on referential expressions.



Q: How many objects are right of the red object?
A: 2.
Q: How many objects have the same material as the cube?
A: 2

III. Interpret complex questions from visual cues.



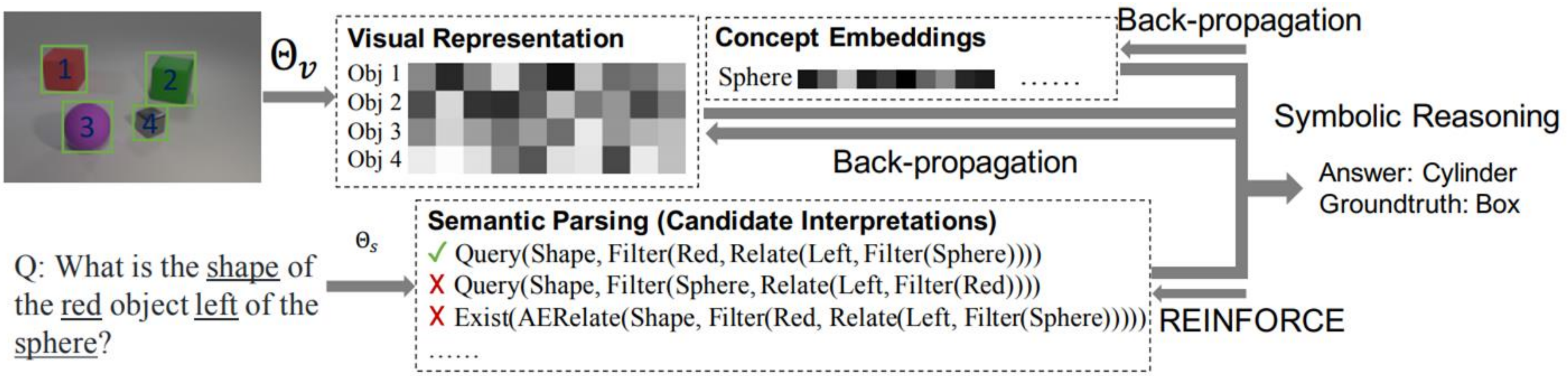
Q: How many objects are both right of the green cylinder and have the same material as the small blue ball?
A: 3

Jiayuan Mao , et al. "The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision." ICLR 2019

The Neuro-symbolic Concept Learner

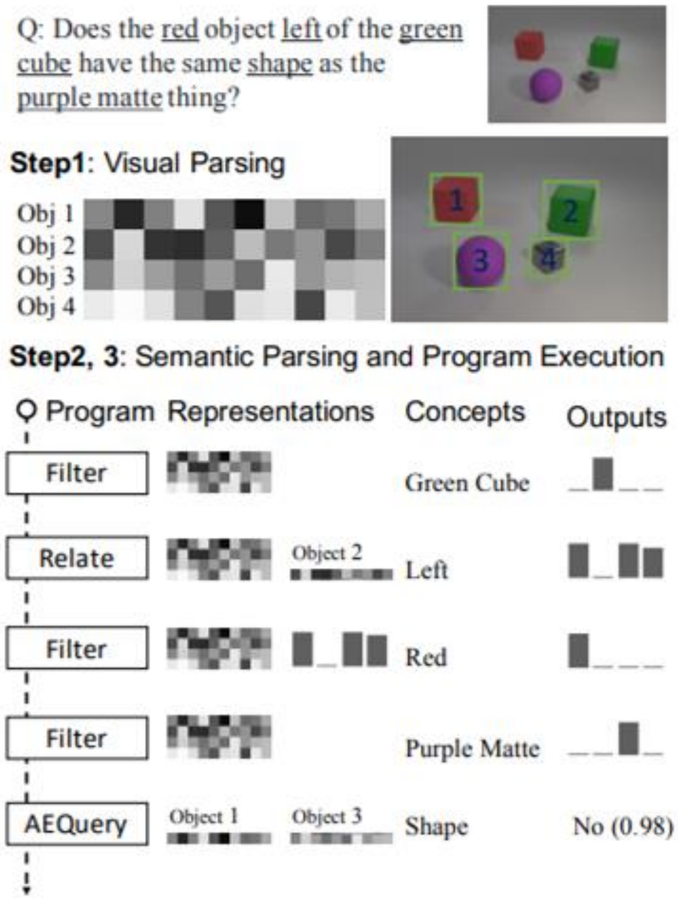
Extension from Neural-symbolic VQA:

Learns visual concepts, words, and semantic parsing of sentences without explicit supervision on any of them, but just by looking at **images and reading paired questions and answers**



Jiayuan Mao , et al. "The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision." ICLR 2019

The Neuro-symbolic Concept Learner

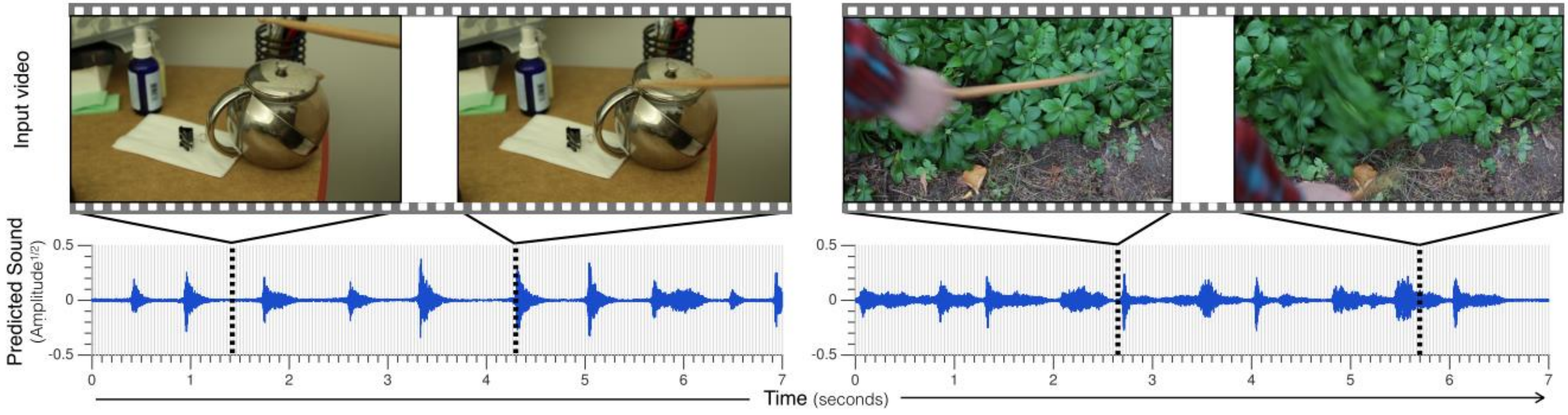


Jiayuan Mao , et al. "The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision." ICLR 2019

Speech-Vision Translation: Applications

Translation 1: Visually indicated sounds

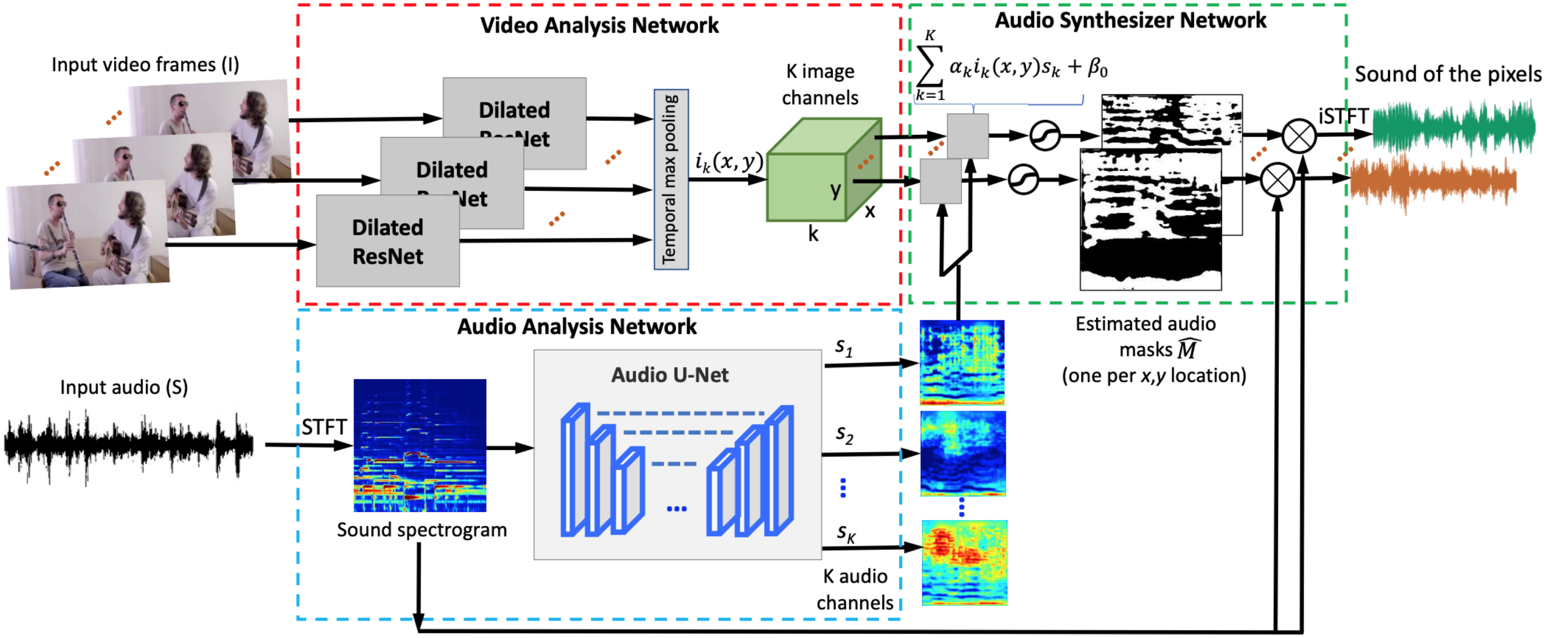
- Sound generation!



[Owens et al. Visually indicated sounds, CVPR, 2016]

Translation 2: The Sound of Pixels

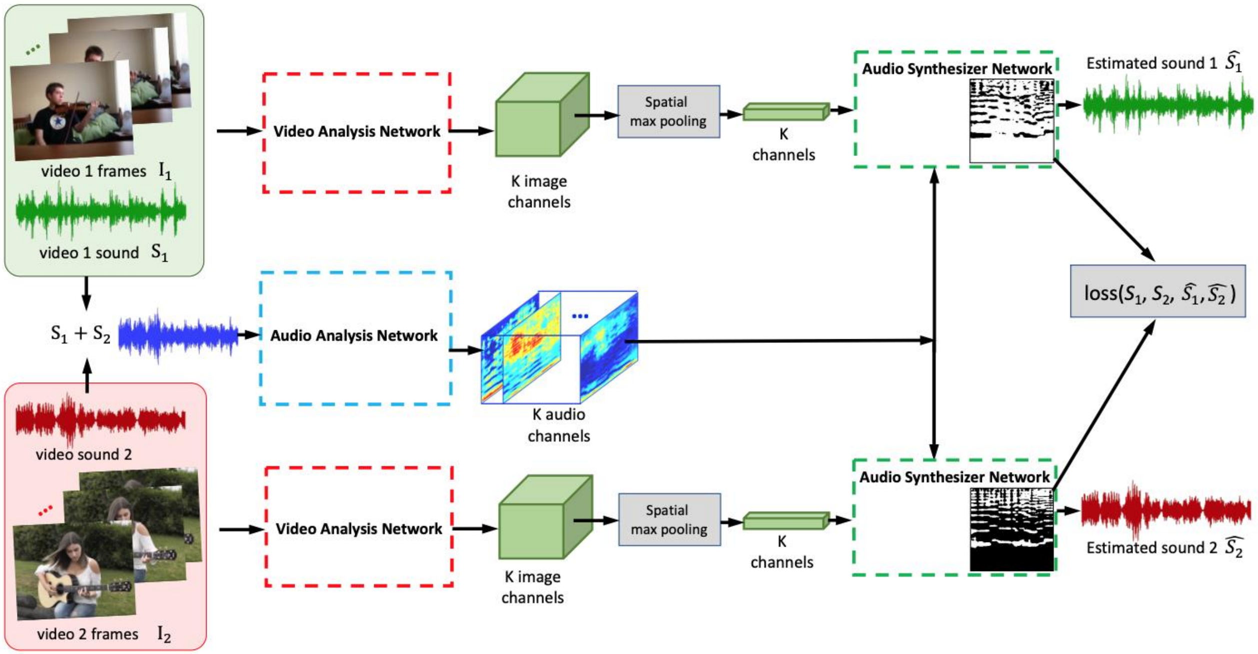
- Propose a system that learns to localize the sound sources in a video and separate the input audio into a set of components coming from each object by leveraging unlabeled videos.



[Zhao, Hang, et al. "The sound of pixels.", ECCV 2018] https://youtu.be/2eVDLEQlKD0

Translation 2: The Sound of Pixels

- Trained in a **self-supervised** manner by learning to separate the sound source of a video from the audio mixture of multiple videos conditioned on the visual input associated with it.

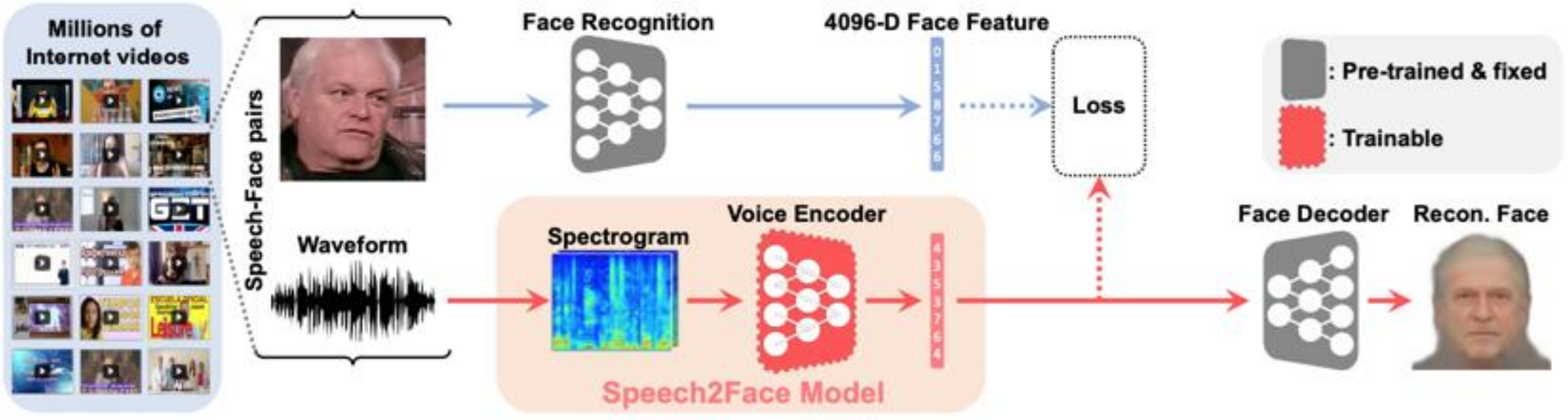


[Zhao, Hang, et al. “The sound of pixels.”, ECCV 2018]

Speech2face

Speech2face

Voice encoder + face encoder + face decoder



Speech2face

Examples of reconstructed faces