Natural Language Processing with Deep Learning
CS224N/Ling284

Lecture 18:
The Limits and Future of NLP

Richard Socher

Poster Session

- Everyone expected to attend (or video).
- 530pm-830pm next Wednesday.
- Check out other student's posters.
- Dinner on us.
- Stay until end if you think you have a chance of winning one of the awards
- Jobs and funding
- Fun :)

What has been lost from old NLP work?

- An earlier era of work had lofty goals, but modest realities
- Today, we have *much* better realities, but often content ourselves with running LSTMs rather than reaching for the stars

Norvig (1986) Ph.D.

Peter Norvig's thesis – 30th anniversary

A Unified Theory of Inference for Text Understanding

By

Peter Norvig

B.S. (Brown University) 1978

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

in the

GRADUATE DIVISION

OF THE

UNIVERSITY OF CALIFORNIA, BERKELEY

Robert Wilensky
Lofti Zadeh
Chuck Fillmore

Approved: 11/24/86
Chairman Date
L. A. Zadeh 11/25/86
Charles J. Fillmore 11/25/86

The language analyzed

- In a poor fishing village built on an island not far from the coast of China, a young boy named Chang Lee lived with his widowed mother. Every day, little Chang bravely set off with his net, hoping to catch a few fish from the sea, which they could sell and have a little money to buy bread.

(a) There is a sea, which surrounds the island, is used by the villagers for fishing, and forms part of the coast of China

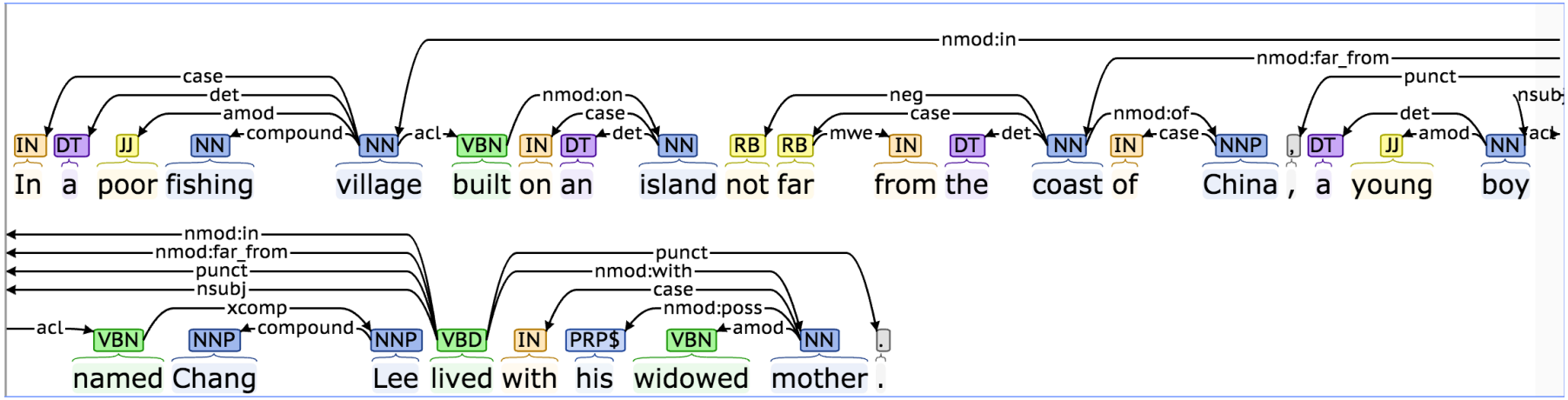
(b) Chang intends to trap fish in his net, which is a fishing net

(c) The word *which* refers to *the fish*

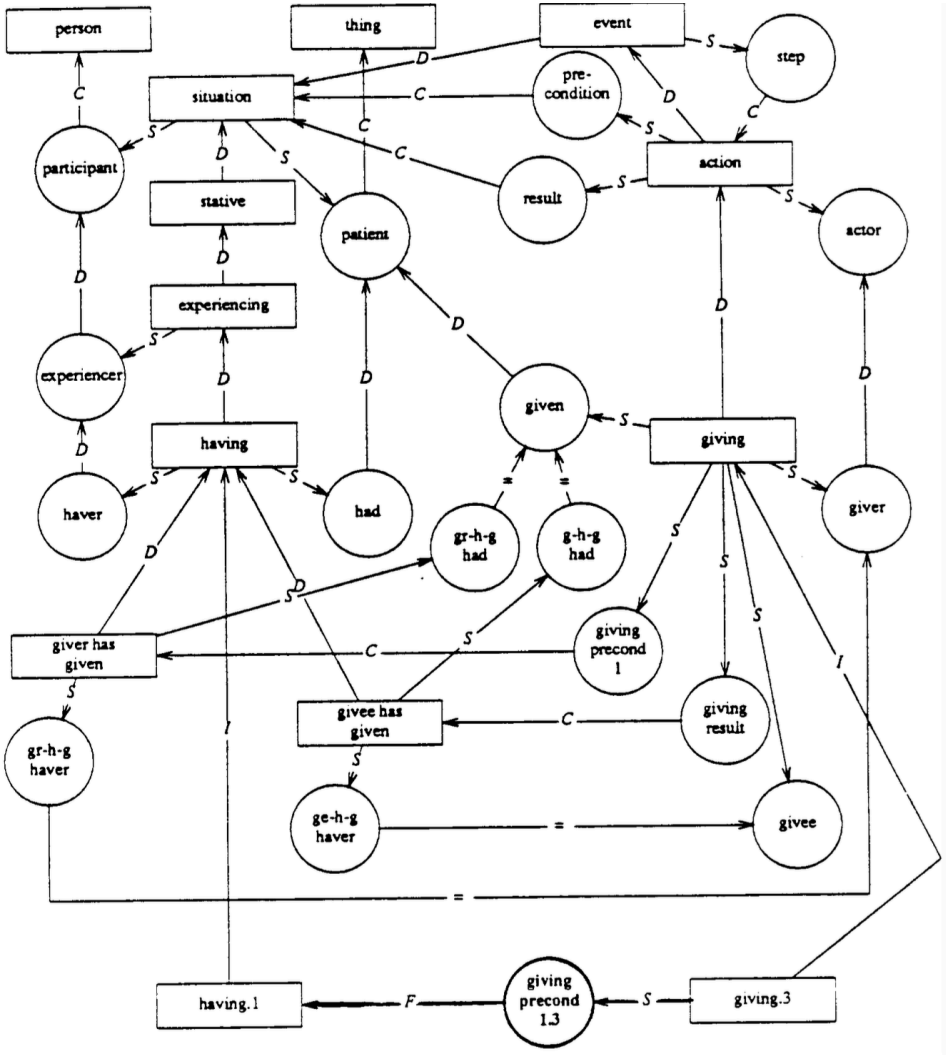
(d) The word *they* refers to Chang and his mother

Basic NLP: Progress has been made!

"Arens and Wilensky's PHRAN program was used where possible [to convert input sentences to KODIAK knowledge representations]. For some input, PHRAN was not up to the task, so a representation was constructed by hand instead." (p. 4)



Building elaborations a la Norvig (1986)

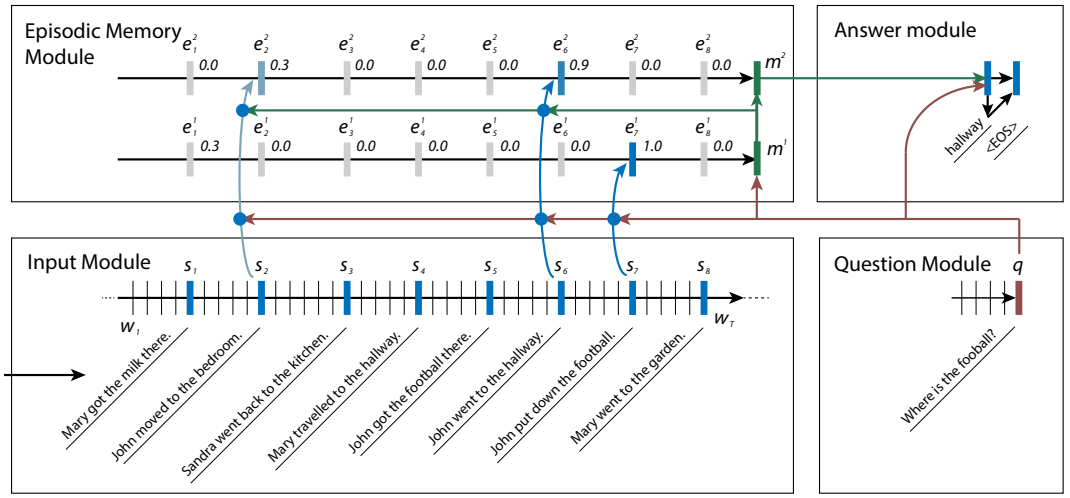


What do we still need?

- BiLSTMs with attention seem to be taking over the field and improving our ability to do **everything**
- Neural methods are leading to a **renaissance** for all language generation tasks (i.e., MT, dialog, QA, summarization, …)
- There's a real scientific question of where and whether we need explicit, localist language and knowledge representations and inferential mechanisms

What do we still need?

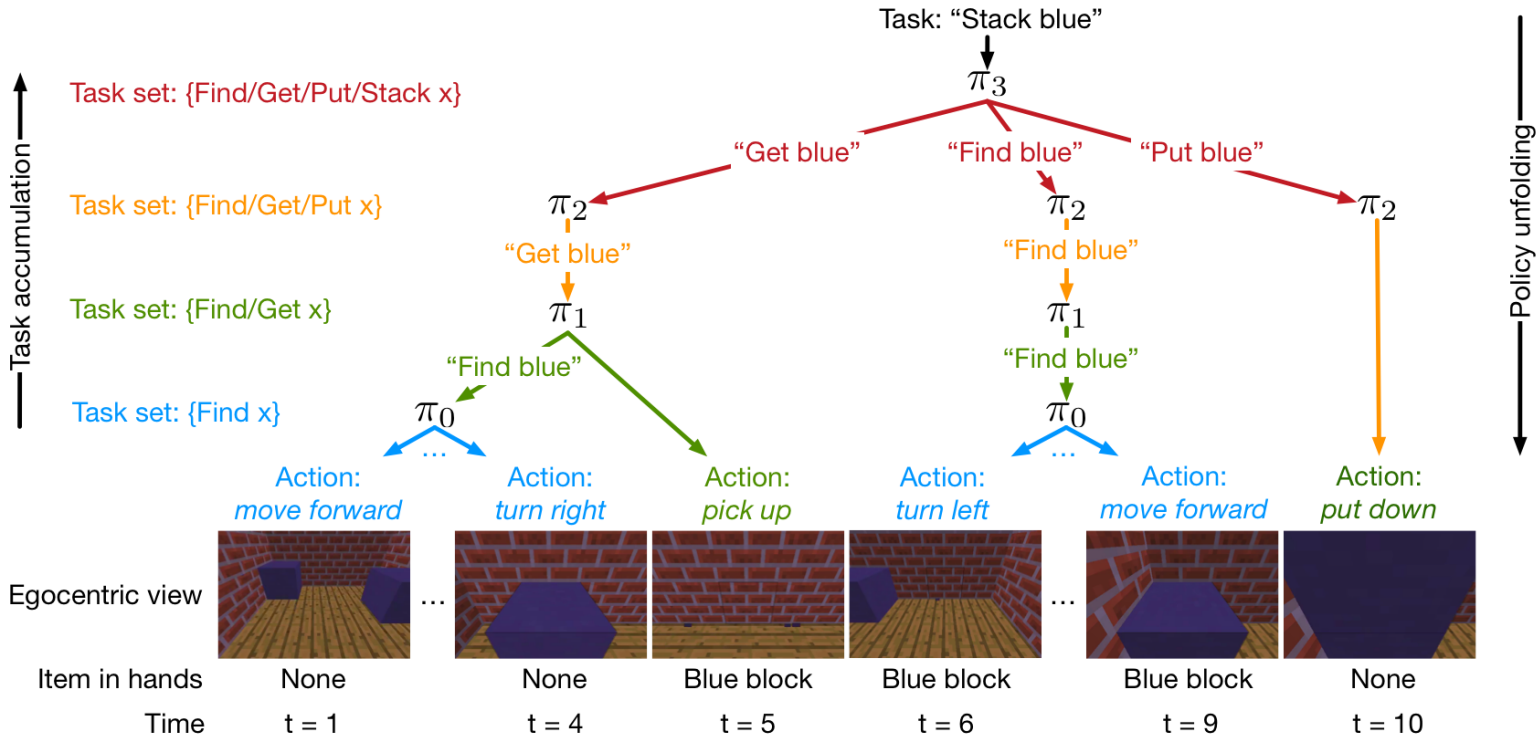
- However: We still have very primitive methods for building and accessing **memories** or **knowledge**



- Current models have almost nothing for developing and executing **goals** and **plans***

Progress on goals and plans

- Hierarchical and Interpretable Skill Acquisition in Multi-task Reinforcement Learning, Tianmin Shu, Caiming Xiong, and Richard Socher International Conference on Learning Representations (**ICLR 2018**)



What do we still need?

- We still have quite inadequate abilities for understanding and using **inter-sentential relationships.**

- We still can't, at a large scale, do **elaborations** from a **situation** using **common sense knowledge BUT also have bias**

The Limits of Single Task Learning

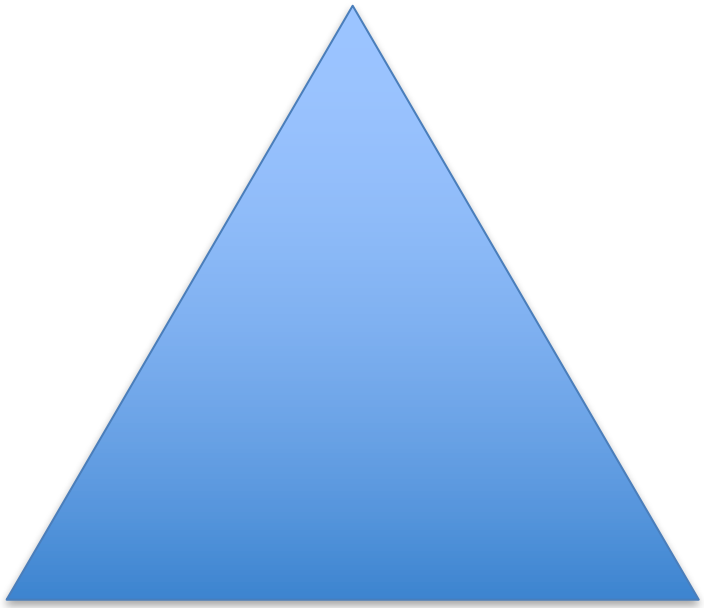
- Great performance improvements
- Projects start from random
- Single unsupervised task can’t fix it
- We will never get to a truly general NLP model this way.

Towards NLP-Complete Super Tasks

- How to express different tasks in the same framework, e.g.
  - Sequence tagging: aspect specific sentiment
  - Text classification: dialogue intent classification
  - Seq2seq: machine translation, summarization, etc.

The 3 Equivalent NLP-Complete Super Tasks

- Language modeling
- Question answering
- Dialogue systems



Usefulness and complexity
in their current interpretation

Framework for Tackling NLP

A joint model for comprehensive QA

QA Examples

I: Mary walked to the bathroom.
I: Sandra went to the garden.
I: Daniel went back to the garden.
I: Sandra took the milk there.
Q: Where is the milk?
A: garden
I: Everybody is happy.
Q: What's the sentiment?
A: positive

I: I think this model is incredible
Q: In French?
A: Je pense que ce modèle est incroyable.

I: 

Q: What color are the bananas?
A: Green.

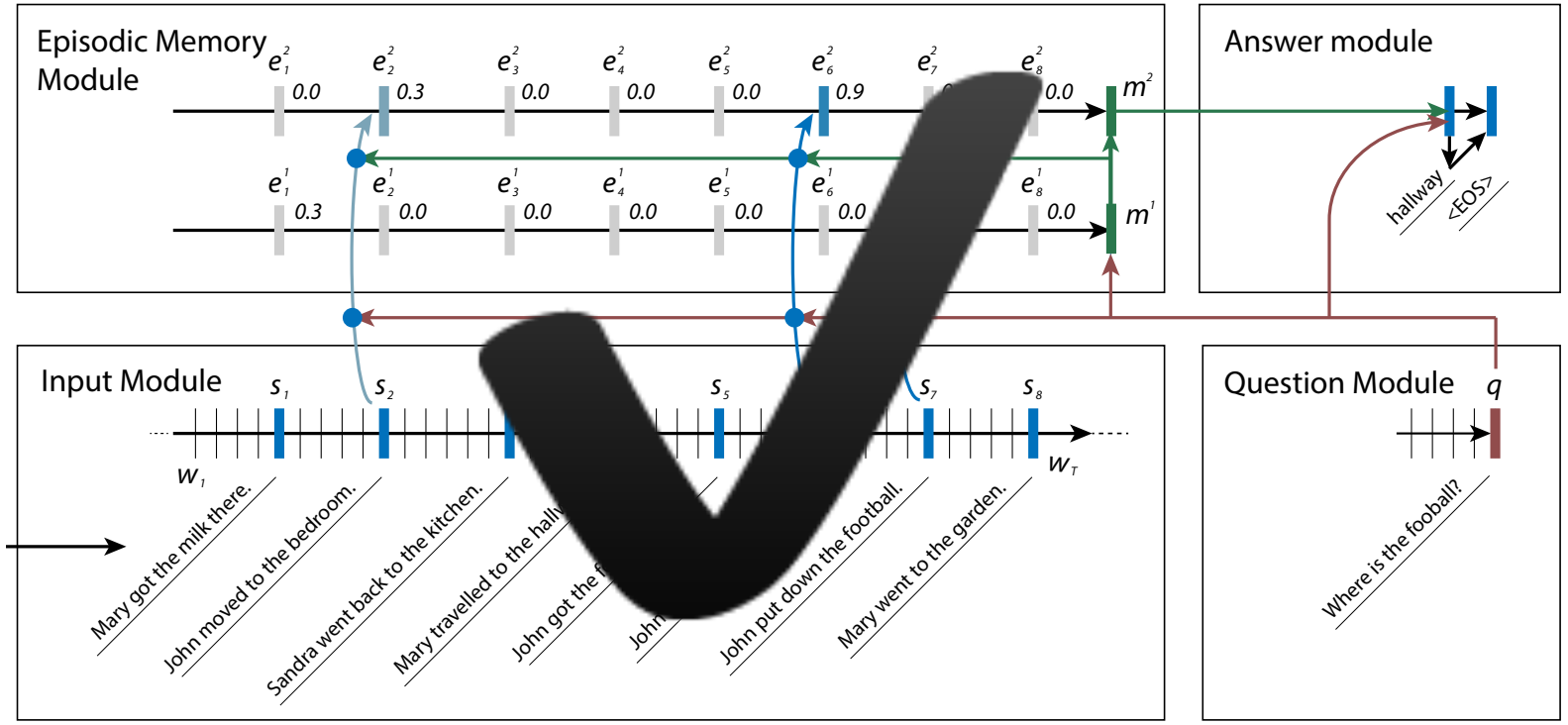
Move from $\{x_i, y_i\}$ to $\{x_i, q_i, y_i\}$

First of Several Major Obstacles

- For NLP no single model **architecture** with consistent state of the art results across tasks

Task	State of the art model
Question answering (babI)	Strongly Supervised MemNN (Weston et al 2015)
Sentiment Analysis (SST)	Tree-LSTMs (Tai et al. 2015)
Part of speech tagging (PTB-WSJ)	Bi-directional LSTM-CRF (Huang et al. 2015)

Tackling Obstacle: Dynamic Memory Network



But, now known, it's not enough

Obstacle: Joint Many-task Learning

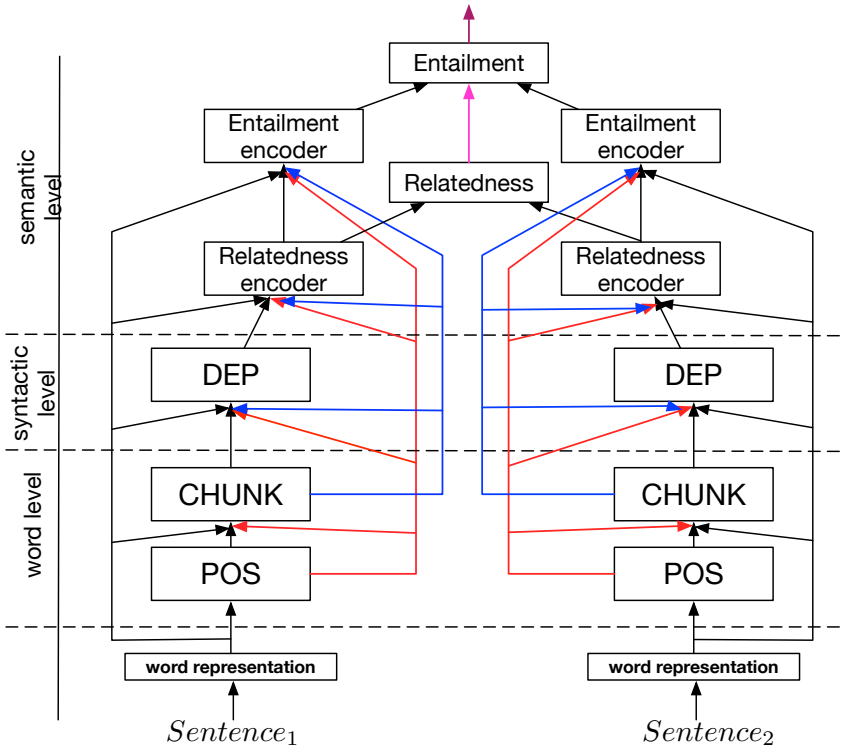
- Fully joint multitask learning* is hard:
  - Usually restricted to lower layers
  - Usually helps only if tasks are related
  - Often hurts performance if tasks are not related
  - We lose powerful accuracy improvement techniques such as task-specific architecture and hyperparameter tuning

* meaning: same decoder/classifier
and not only transfer learning with source target task pairs, no swappable modeling blocks per task

Tackling Joint Training

- A Joint Many-Task Model:
  Growing a Neural Network for Multiple NLP Tasks
  Kazuma Hashimoto,
  Caiming Xiong,
  Yoshimasa Tsuruoka &
  Richard Socher

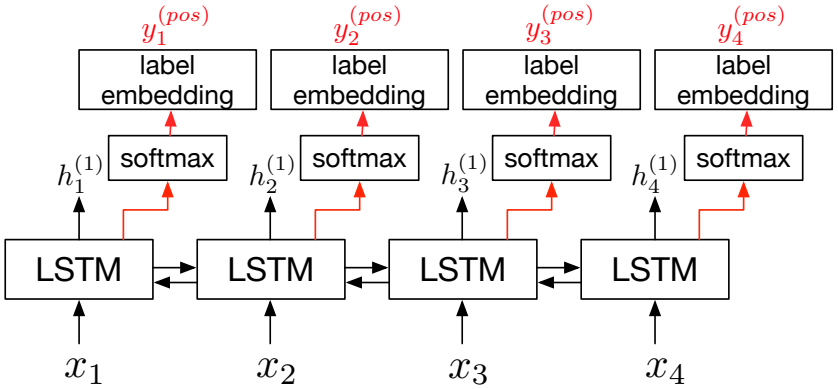
- Final Model →



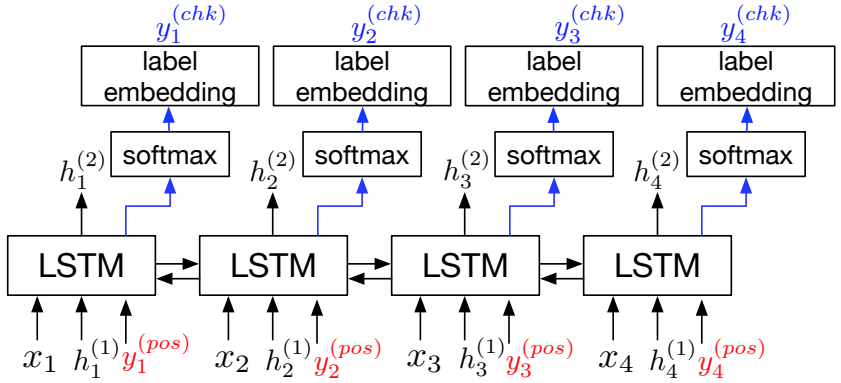
Model Details

- Include character n-grams and short-circuits
- State of the art purely feedforward parser

POS Tagging:

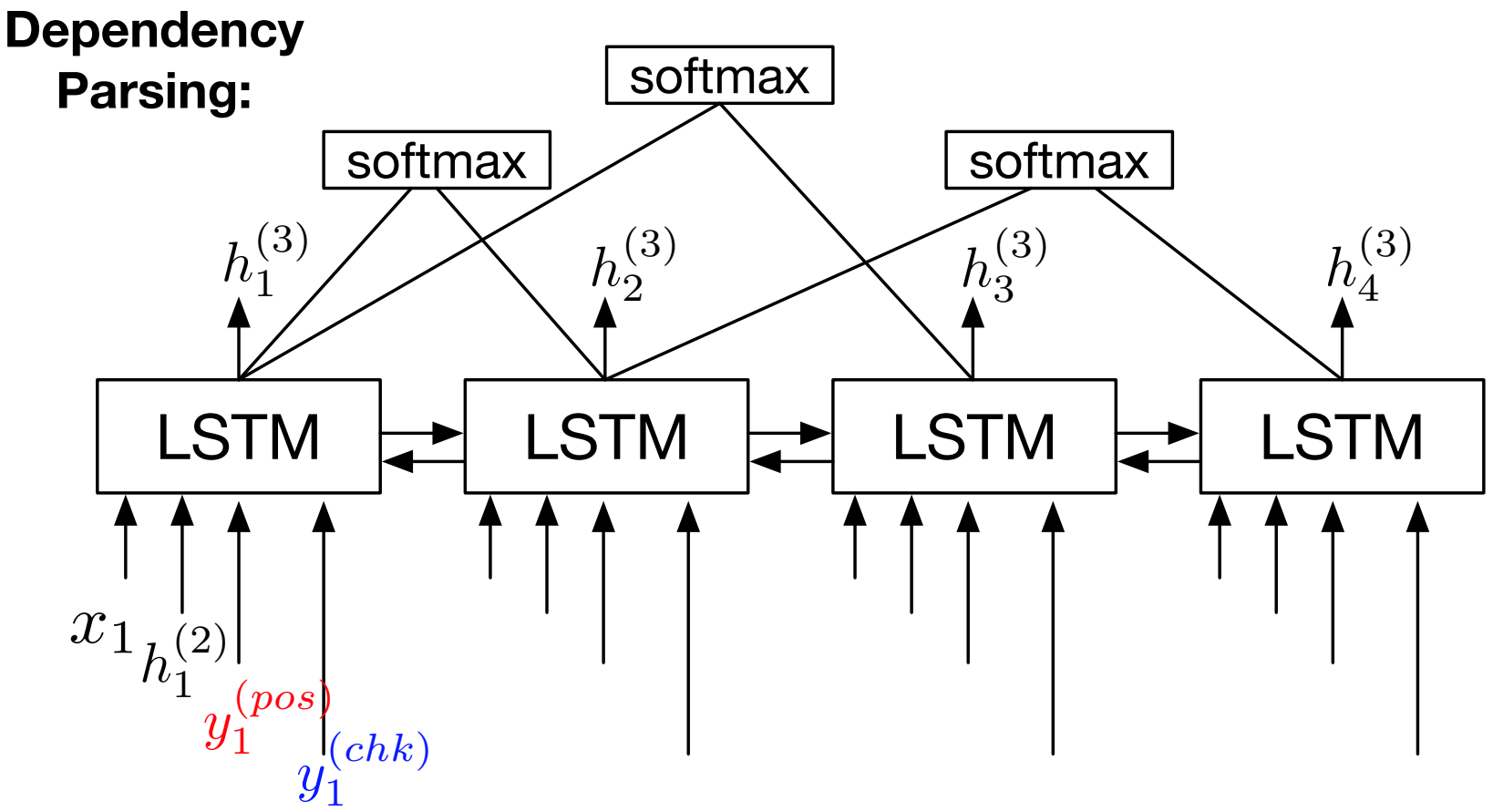


Chunking:

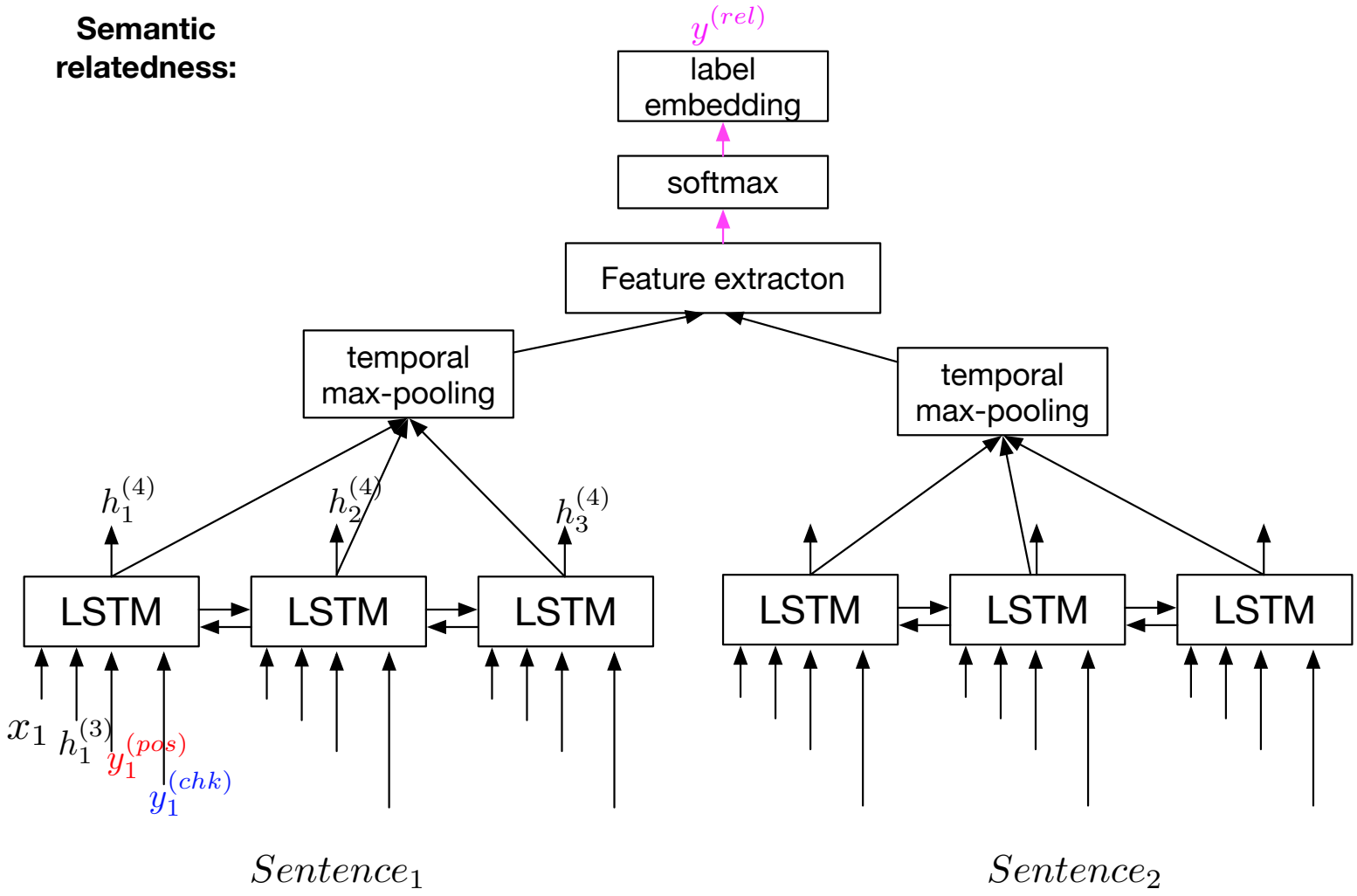


$$y_t^{(pos)} = \sum_{j=1}^{C} p(y_t^{(1)} = j | h_t^{(1)}) \ell(j),$$

Dependency Parsing



Multi Sentence Tasks: Semantic Relatedness



Training Details: Regularized Idea

Chunking training

$$-\sum_{s}\sum_{t}\log p(y_t^{(2)}=\alpha|h_t^{(2)})+\lambda\|W_{\text{chunk}}\|^2+\delta\|\theta_{\text{POS}}-\theta'_{\text{POS}}\|^2,$$

Entailment training

$$-\sum_{(s,s')}\log p(y_{(s,s')}^{(5)}=\alpha|h_s^{(5)},h_{s'}^{(5)})+\lambda\|W_{\text{ent}}\|^2+\delta\|\theta_{\text{rel}}-\theta'_{\text{rel}}\|^2,$$

Joint Training Helps Here!

		Single	JMT_{all}	JMT_{AB}	JMT_{ABC}	JMT_{DE}
A	POS	97.45	97.55	97.52	97.54	n/a
B	Chunking	95.02	(97.12)	95.77	(97.28)	n/a
C	Dependency UAS	93.35	94.67	n/a	94.71	n/a
	Dependency LAS	91.42	92.90	n/a	92.92	n/a
D	Relatedness	0.247	0.233	n/a	n/a	0.238
E	Entailment	81.8	86.2	n/a	n/a	86.8

New State of the Art on 4 of 5 Tasks

Method	Acc.
JMT_{all}	97.55
Ling et al. (2015)	**97.78**
Kumar et al. (2016)	97.56
Ma & Hovy (2016)	97.55
Søgaard (2011)	97.50
Collobert et al. (2011)	97.29
Tsuruoka et al. (2011)	97.28
Toutanova et al. (2003)	97.27

Table 2: POS tagging results.

Method	F1
JMT_{AB}	**95.77**
Søgaard & Goldberg (2016)	95.56
Suzuki & Isozaki (2008)	95.15
Collobert et al. (2011)	94.32
Kudo & Matsumoto (2001)	93.91
Tsuruoka et al. (2011)	93.81

Table 3: Chunking results.

Method	UAS	LAS
JMT_{all}	**94.67**	**92.90**
Single	93.35	91.42
Andor et al. (2016)	94.61	92.79
Alberti et al. (2015)	94.23	92.36
Weiss et al. (2015)	93.99	92.05
Dyer et al. (2015)	93.10	90.90
Bohnet (2010)	92.88	90.71

Table 4: Dependency results.

Method	MSE
JMT_{all}	**0.233**
JMT_{DE}	0.238
Zhou et al. (2016)	0.243
Tai et al. (2015)	0.253

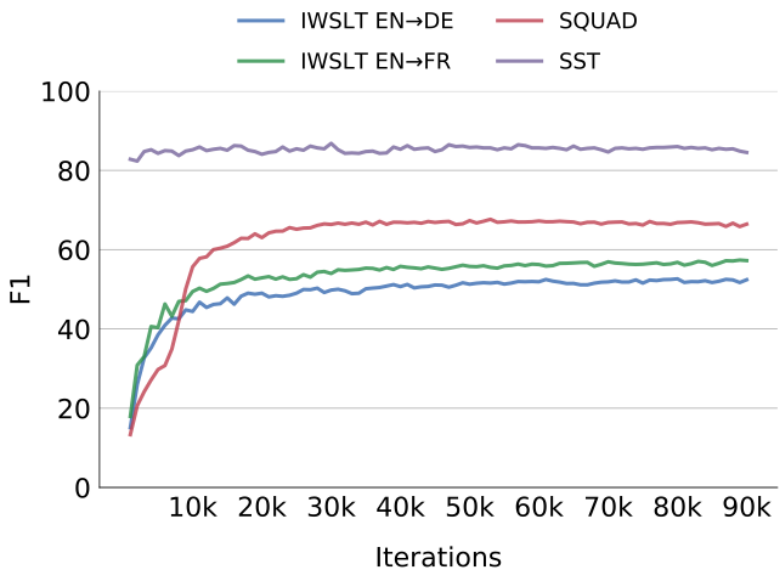
Table 5: Semantic relatedness results.

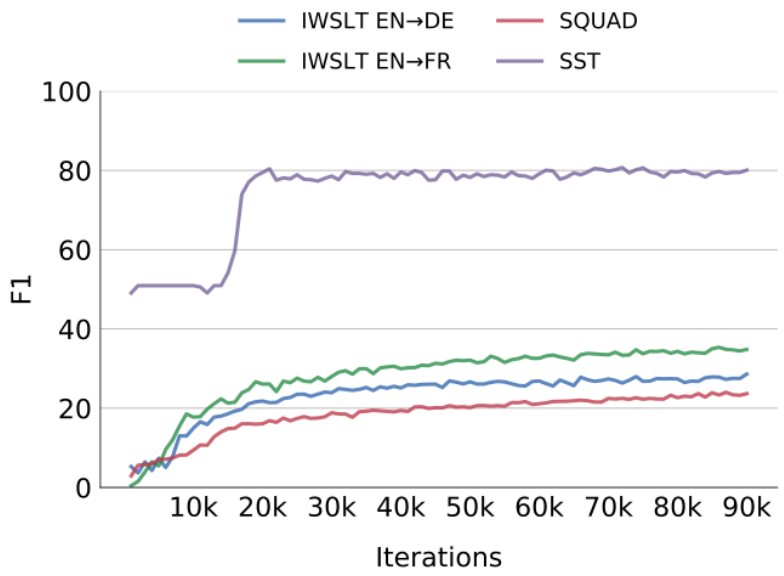
Method	Acc.
JMT_{all}	86.2
JMT_{DE}	**86.8**
Yin et al. (2016)	86.2
Lai & Hockenmaier (2014)	84.6

Table 6: Textual entailment results.

Progress of Recent Weeks

- Joint model trained on harder tasks
- Single task models Joint many-task model

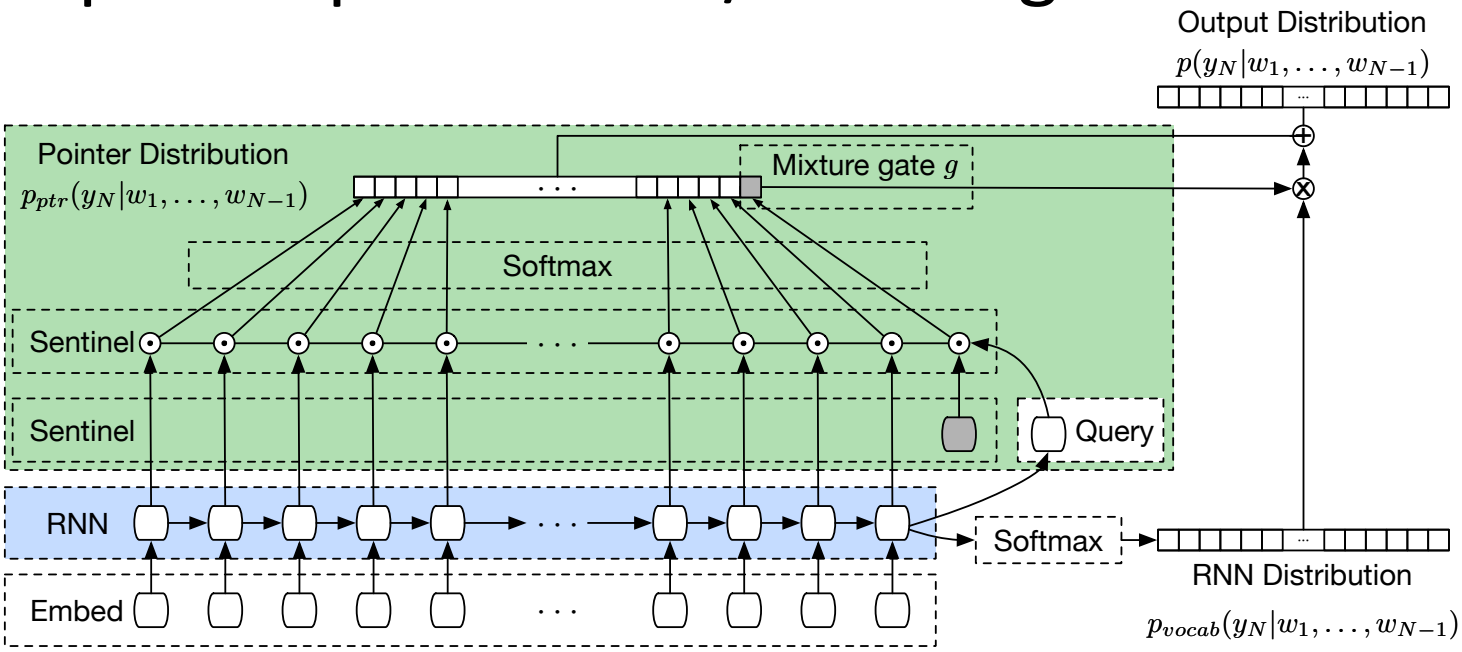




- First solution described in class

Obstacle: Duplicate Word Representations

- Different encodings for encoder (Word2Vec and GloVe word vectors) and decoder (softmax classification weights for words)
- Duplicate parameters/meaning



Tackling Obstacle by Tying Word Vectors

- Simple but theoretically motivated idea: tie word vectors and train single weights jointly

- Tying Word Vectors and Word Classifiers: A Loss Framework for Language Modeling, Hakan Inan, Khashayar Khosravi, Richard Socher (ICLR 2017)

Language Modeling With Tying Word Vectors

MODEL	PARAMETERS	VALIDATION	TEST
KN-5 (Mikolov & Zweig)	2M	-	141.2
KN-5 + Cache (Mikolov & Zweig)	2M	-	125.7
RNN (Mikolov & Zweig)	6M	-	124.7
RNN+LDA (Mikolov & Zweig)	7M	-	113.7
RNN+LDA+KN-5+Cache (Mikolov & Zweig)	9M	-	92.0
Deep RNN (Pascanu et al., 2013a)	6M	-	107.5
Sum-Prod Net (Cheng et al., 2014)	5M	-	100.0
LSTM (medium) (Zaremba et al., 2014)	20M	86.2	82.7
LSTM (large) (Zaremba et al., 2014)	66M	82.2	78.4
VD-LSTM (medium, untied) (Gal, 2015)	20M	81.9 ± 0.2	79.7 ± 0.1
VD-LSTM (medium, untied, MC) (Gal, 2015)	20M	-	78.6 ± 0.1
VD-LSTM (large, untied) (Gal, 2015)	66M	77.9 ± 0.3	75.2 ± 0.2
VD-LSTM (large, untied, MC) (Gal, 2015)	66M	-	73.4 ± 0.0
CharCNN (Kim et al., 2015)	19M	-	78.9
VD-RHN (Zilly et al., 2016)	32M	72.8	71.3
Pointer Sentinel-LSTM(medium) (Merity et al., 2016)	21M	72.4	70.9
38 Large LSTMs (Zaremba et al., 2014)	2.51B	71.9	68.7
10 Large VD-LSTMs (Gal, 2015)	660M	-	68.7
VD-LSTM +REAL (medium)	14M	75.7	73.2
VD-LSTM +REAL (large)	51M	**71.1**	**68.5**

Obstacle: Necessary Inputs to QA

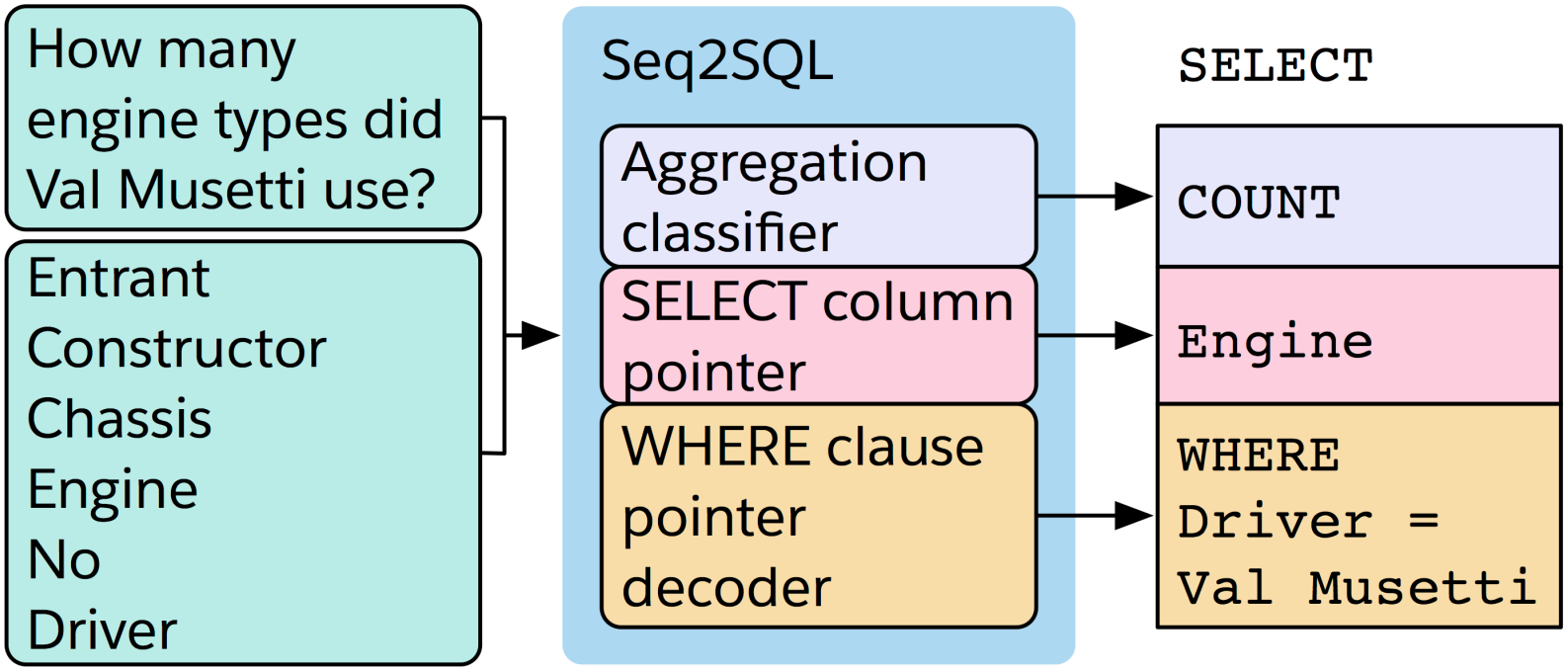
- We need to be able to understand text, **images and databases** to really answer all kinds of questions

Database QA

- **Question:** Who was drafted with the 3rd pick of the 1st round?
- **Answer:** Jayson Tatum

Rnd	Pick	Player	Pos	Nationality	Team	School/club
1	1	Markelle Fultz	PG	United States	Philadelphia 76ers (from Brooklyn via Boston)	Washington
1	2	Lonzo Ball	PG	United States	Los Angeles Lakers	UCLA
1	3	Jayson Tatum	SF	United States	Boston Celtics (from Sacramento via Philadelphia)	Duke
1	4	Josh Jackson	SF	United States	Phoenix Suns	Kansas
1	5	De'Aaron Fox	PG	United States	Sacramento Kings	Kentucky
1	6	Jonathan Isaac	SF/PF	United States	Orlando Magic	Florida State
1	7	Lauri Markkanen	PF/C	Finland	Minnesota Timberwolves (traded to Chicago Bulls)	Arizona

Seq2SQL Overview



Seq2SQL for DB QA

Database QA Results

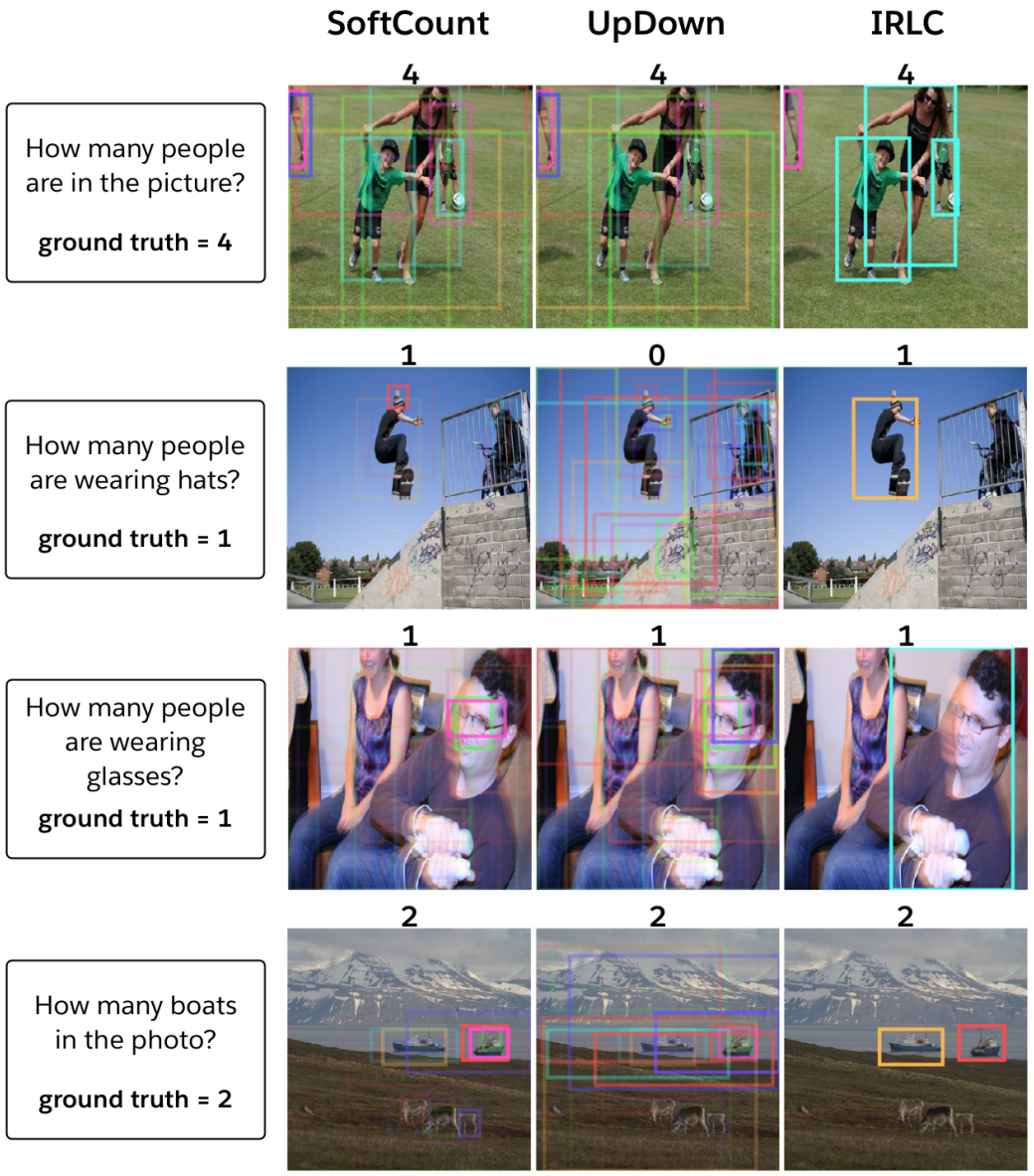
Model	Test Logical Form Accuracy	Test Execution Accuracy
Attentional Seq2Seq	23.4	35.9
Augmented Pointer Network	42.8	52.8
Seq2SQL (no RL)	47.2	57.6
Seq2SQL	**49.2**	**60.3**

Recent Visual Question Answering

- Interpretable Reinforcement Learning Counter
- Good for discrete reasoning over images
- Interpretable Counting for Visual Question Answering, Alexander Trott, Caiming Xiong, Richard Socher. **ICLR 2018**

How many people are pictured?

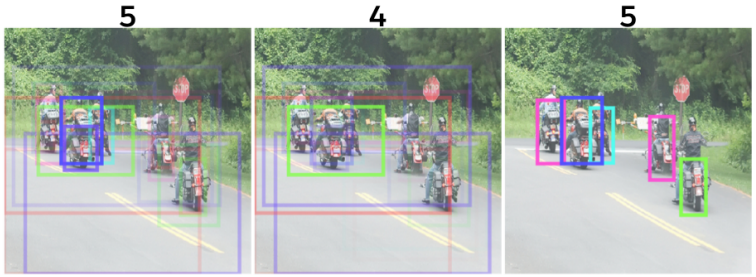
Results


SoftCount
UpDown
IRLC
How many people are in the picture?
ground truth = 4
4
4
4
How many people are wearing hats?
ground truth = 1
1
0
1
How many people are wearing glasses?
ground truth = 1
1
1
1
How many boats in the photo?
ground truth = 2
2
2
2

Results

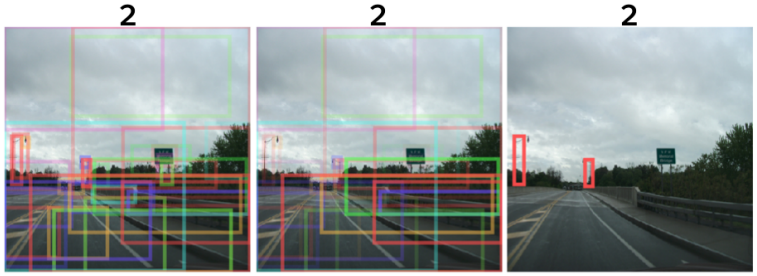
How many motorcycles are in the picture?

ground truth = 5



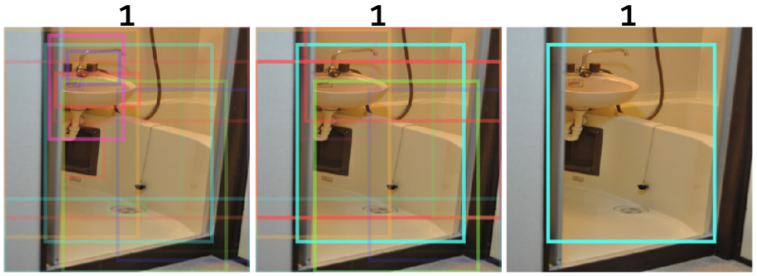
How many utility light poles are pictured?

ground truth = 2



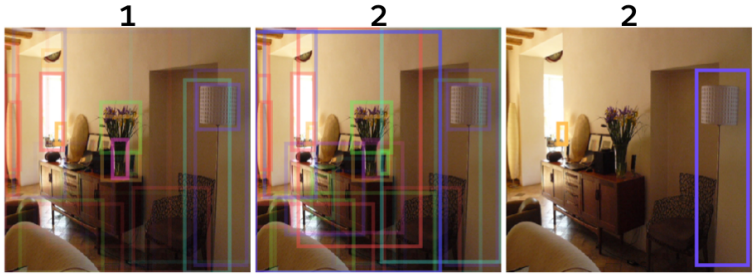
How many drains are visible in the picture?

ground truth = 1



How many lamps are there?

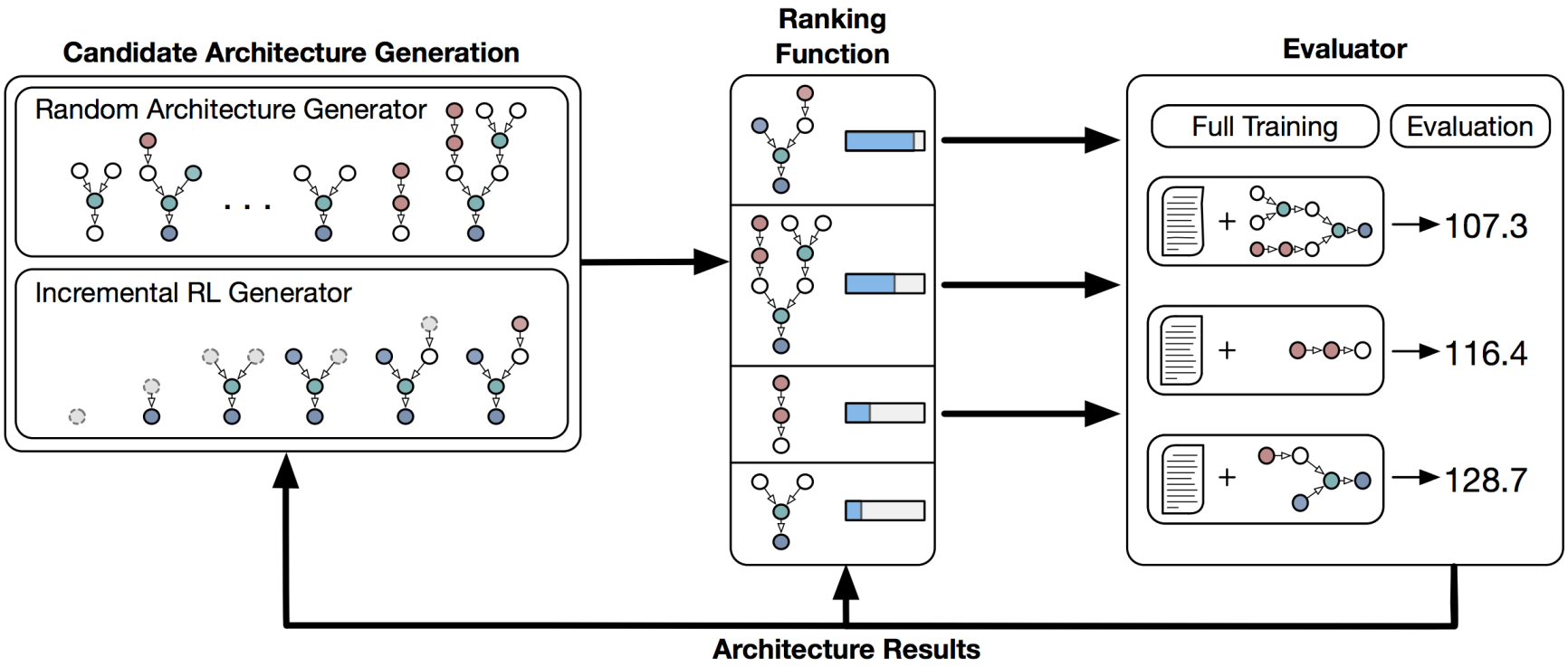
ground truth = 1



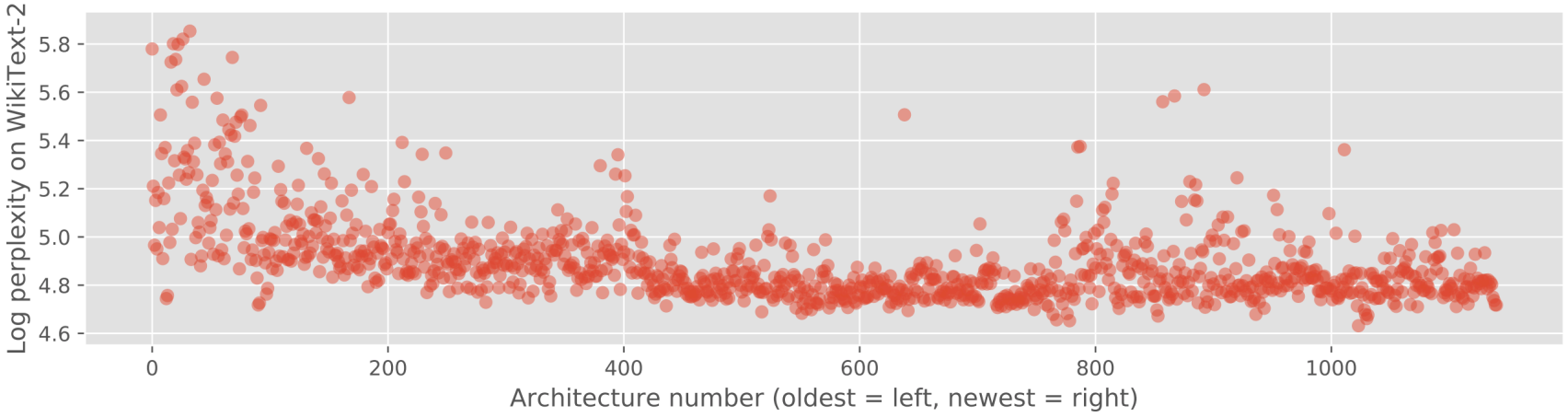
Obstacle: Architecture Engineering

- We don’t yet know the right model architecture for comprehensive QA & joint multitask learning
- Architecture Search is an active area of research but usually applied to simpler/known tasks, e.g.
- A Flexible Approach to Automated RNN Architecture Generation, Stephen Merity, Martin Schrimpf, James Bradbury, Richard Socher. (ICLR 2018 Workshop Track)

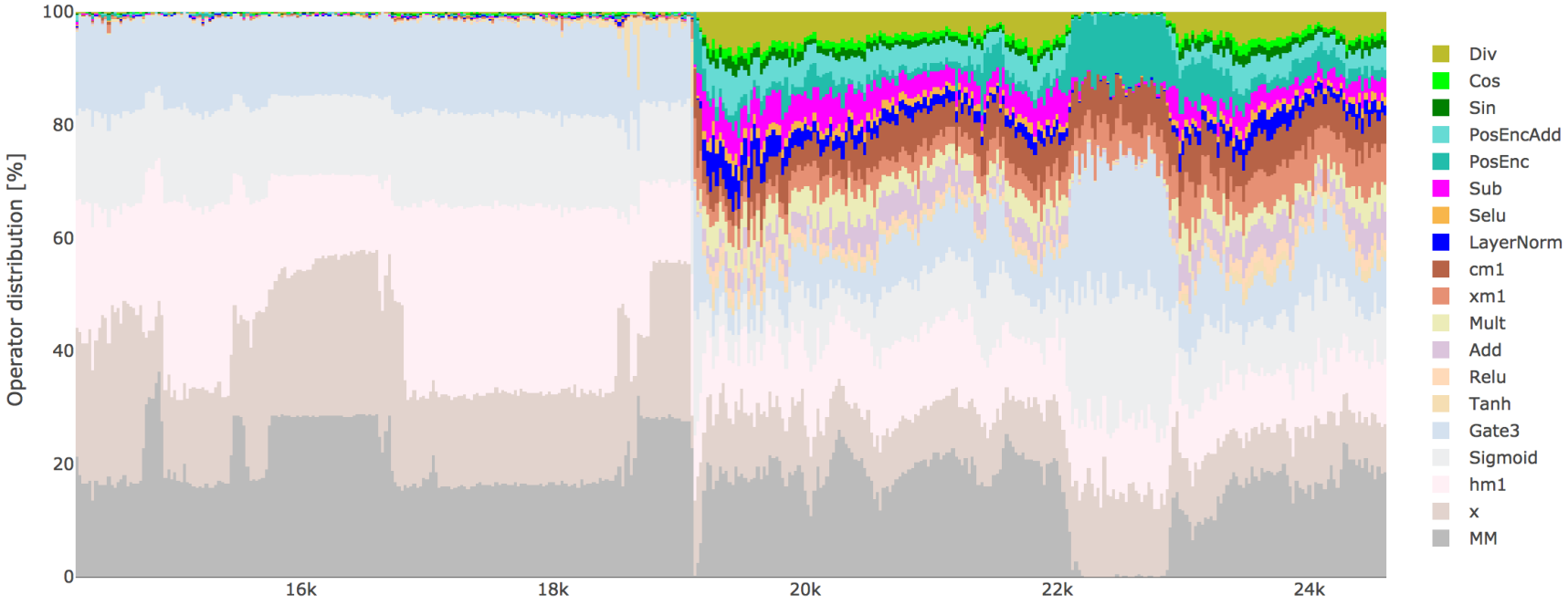
Architecture Search Overview



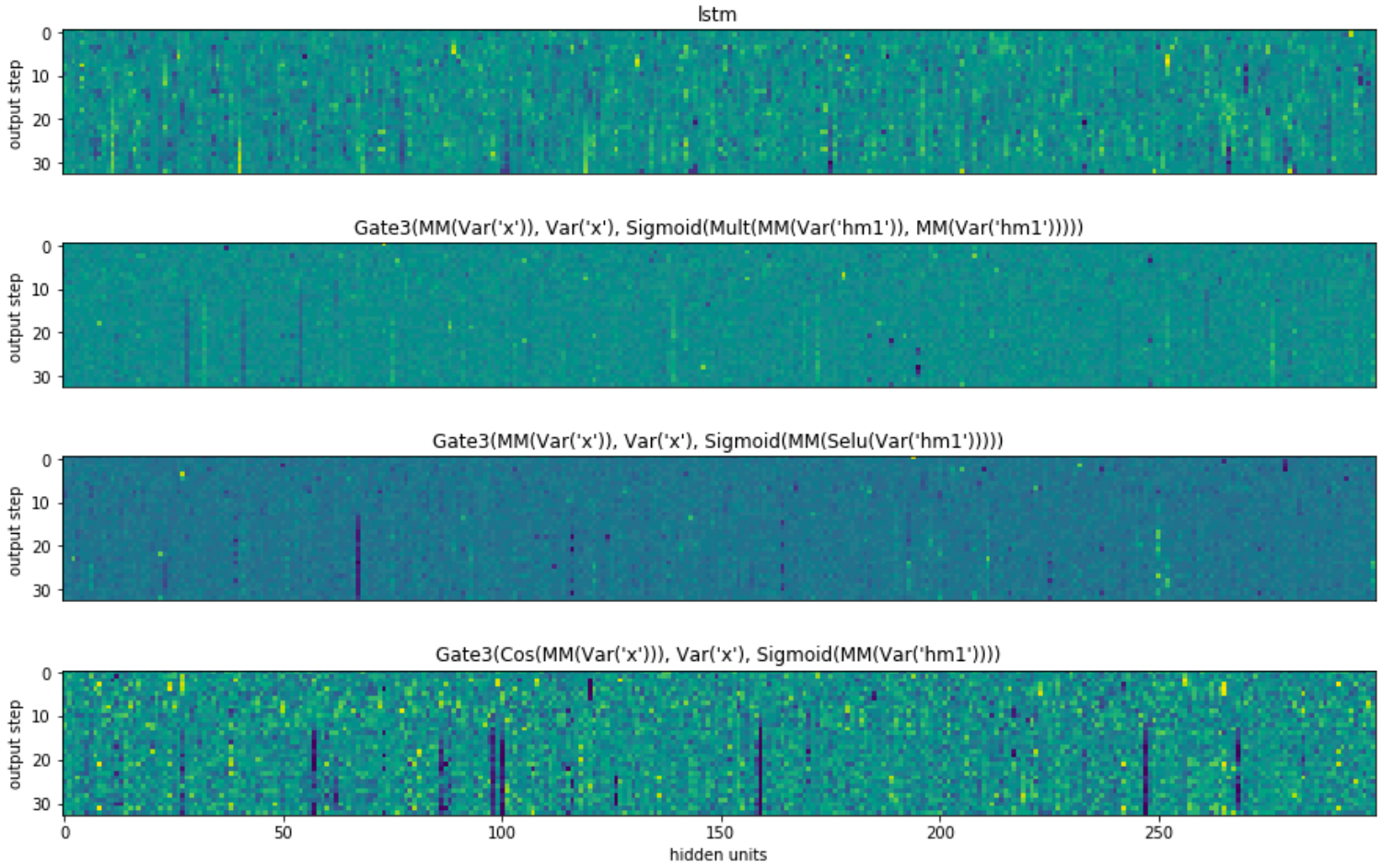
When trained on language modeling



Interesting new building blocks like cos



Very different activation patterns!

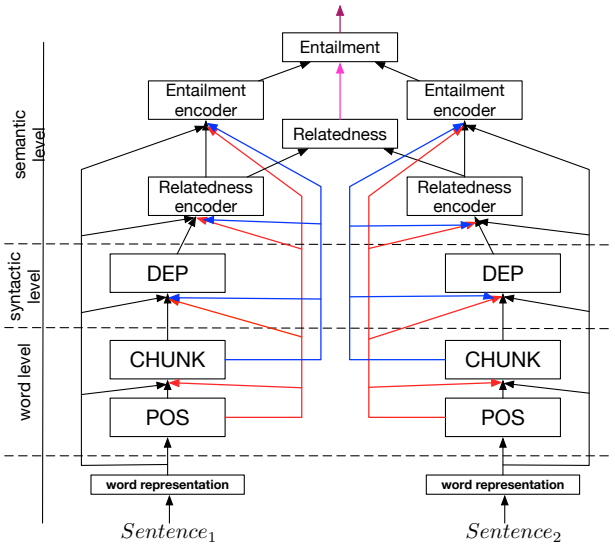


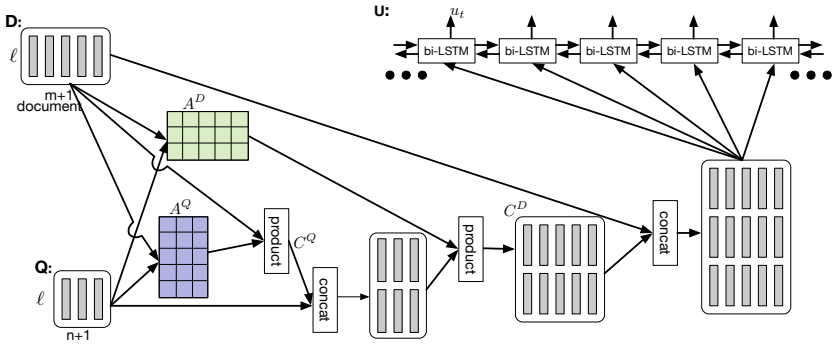
Recent Work on Architecture Search

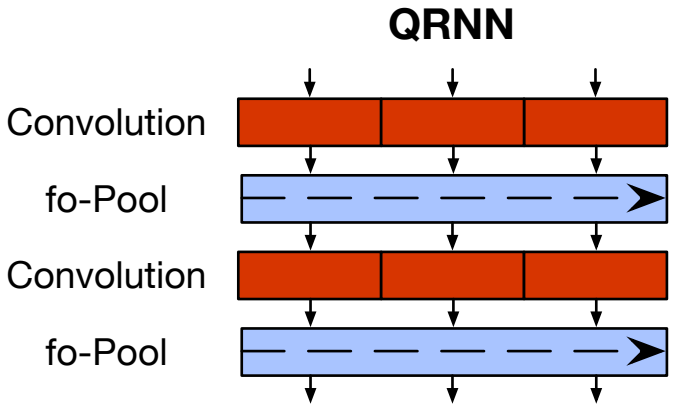
- Efficient Neural Architecture Search via Parameter Sharing
- Hieu Pham, Melody Y. Guan, Barret Zoph, Quoc V. Le, Jeff Dean
- Mon, 12 Feb 2018 :)
- **1000x more efficient and finds better models**
- Shares parameters between models instead of training from scratch

Lots of Limits for deepNLP

- Comprehensive QA
- Multitask learning
- Combined multimodal, logical and memory-based reasoning
- Learning from few examples







DeepNLP

Congratulations!

Good luck with the projects