Fall 2020

TEXAS
The University of Texas at Austin

CS 395T: Topics in Natural Language Processing

11/03/2020

QUESTION ANSWERING AND REASONING

Overview

Span Extraction

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?

Model - BIDAF

Reasoning

Denver would retake the lead with kicker **Matt Prater nailing a 43-yard field goal**, yet Carolina answered as kicker **John Kasay ties the game with a 39-yard field goal**. ... Carolina closed out the half with **Kasay nailing a 44-yard field goal**. ... In the fourth quarter, Carolina sealed the win with **Kasay's 42-yard field goal**.	Which kicker kicked the most field goals?

Model – Neural Module Network

FALL 2020

TEXAS
The University of Texas at Austin

BI-DIRECTIONAL ATTENTION FLOW FOR MACHINE COMPREHENSION

Minjoon Seo[1], Aniruddha Kembhavi[2], Ali Farhadi[1,2], Hananneh Hajishirzi[1]
University of Washington[1], Allen Institute for Artificial Intelligence

SUNDARA RAMAN RAMACHANDRAN
The University of Texas at Austin

TASK

- Machine Comprehension
  - Answering a query about a given context paragraph
  - Two Datasets
    - SQuAD 1.1
      - Answer must be a single span
      - Answer is always a subphrase in the paragraph
    - CNN and Dailymail
      - Answer is one word

DATASETS

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Figure 1: Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

SQuAD 1.1

https://arxiv.org/abs/1606.05250

Context

the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " . . .

Question

producer **X** will not press charges against *ent212* , his lawyer says .

Answer

ent193

Daily Mail

https://arxiv.org/abs/1506.03340

Attention Mechanisms in Prior works to BIDAF:

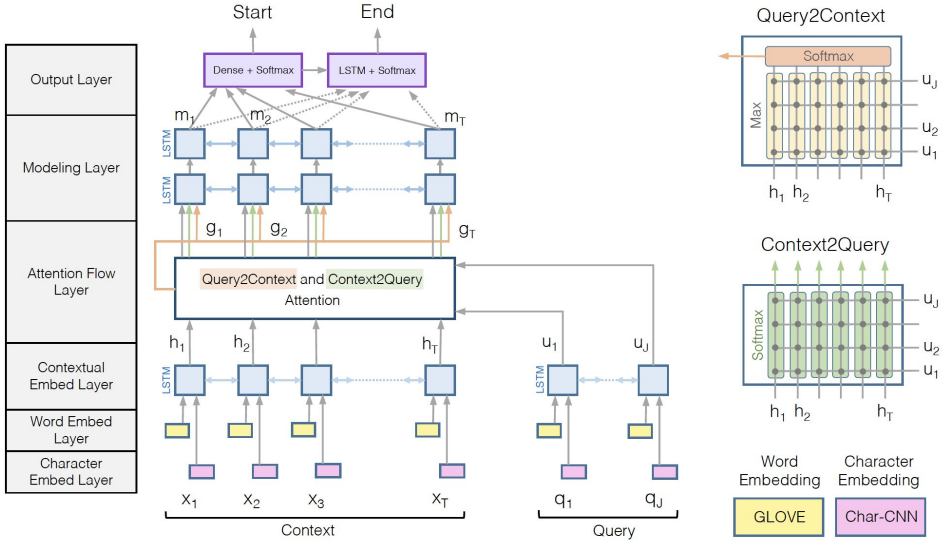
1. Context is summarized into a fixed size vector
2. Temporally Dynamic
3. Uni – Directional
   - Query to Context

BIDAF

1. Context paragraph is not summarized into a fixed size vector.
2. Memory less attention mechanism
3. Bi-directional Attention
   - Query2Context
   - Context2Query

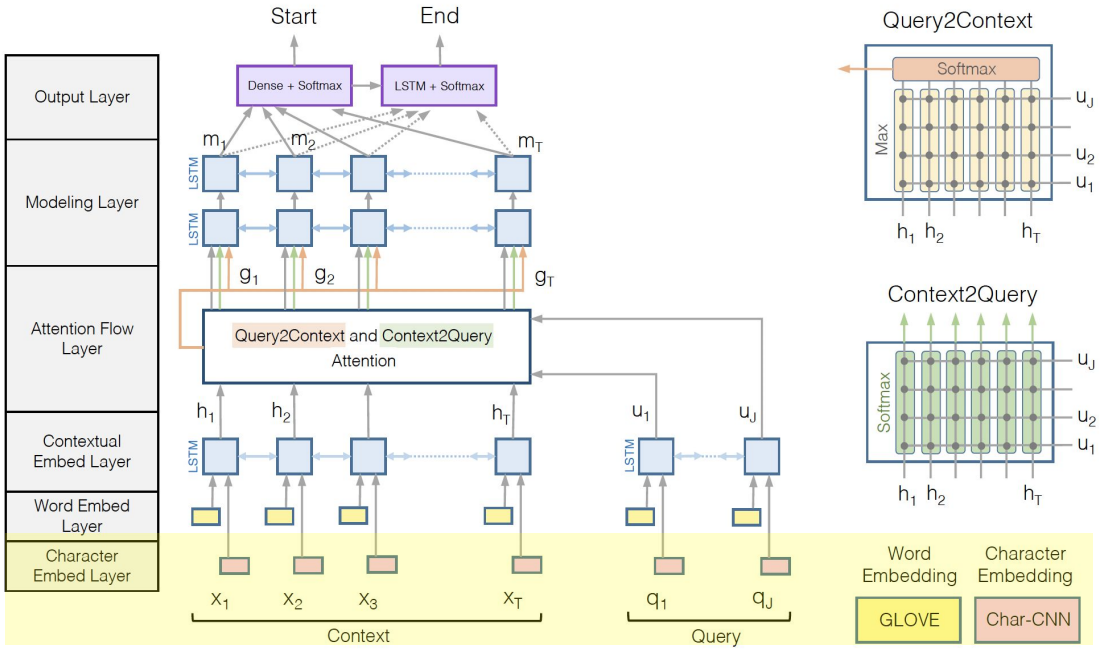
❖ Note: This work is before ELMO, BERT, etc., but after GLoVE

MODEL


Output Layer
Modeling Layer
Attention Flow Layer
Contextual Embed Layer
Word Embed Layer
Character Embed Layer
Start
End
Dense + Softmax
LSTM + Softmax
LSTM
Query2Context and Context2Query Attention
Context
Query
Query2Context
Softmax
Max
Context2Query
Word Embedding
Character Embedding
GLOVE
Char-CNN

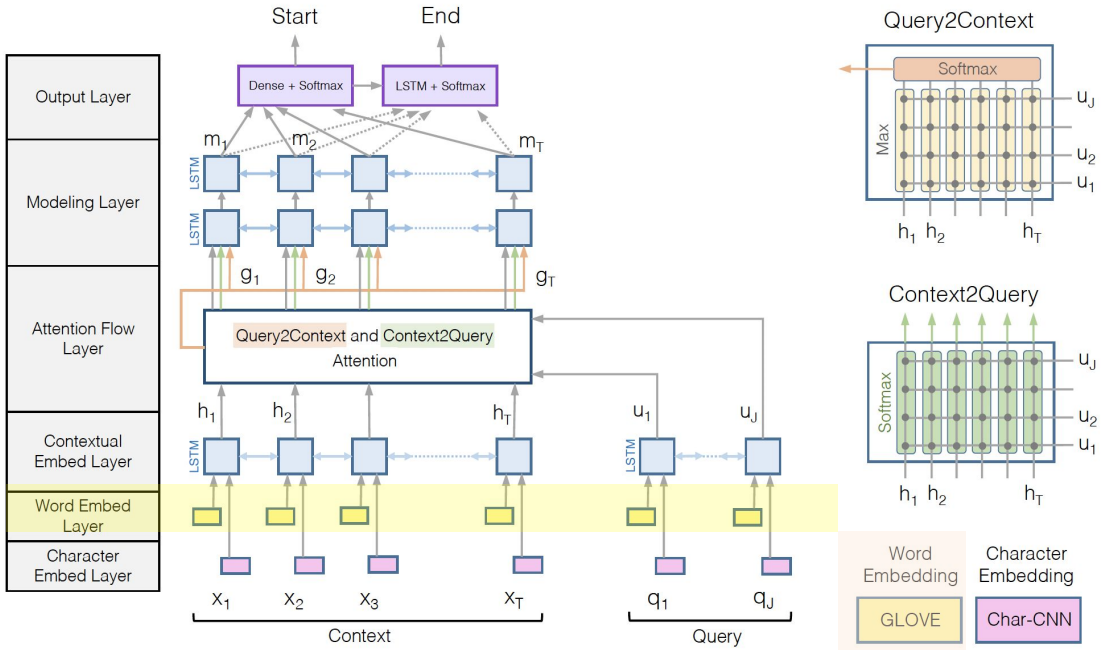
Character Embedding Layer

Maps each word to a vector space using character-level CNNs.



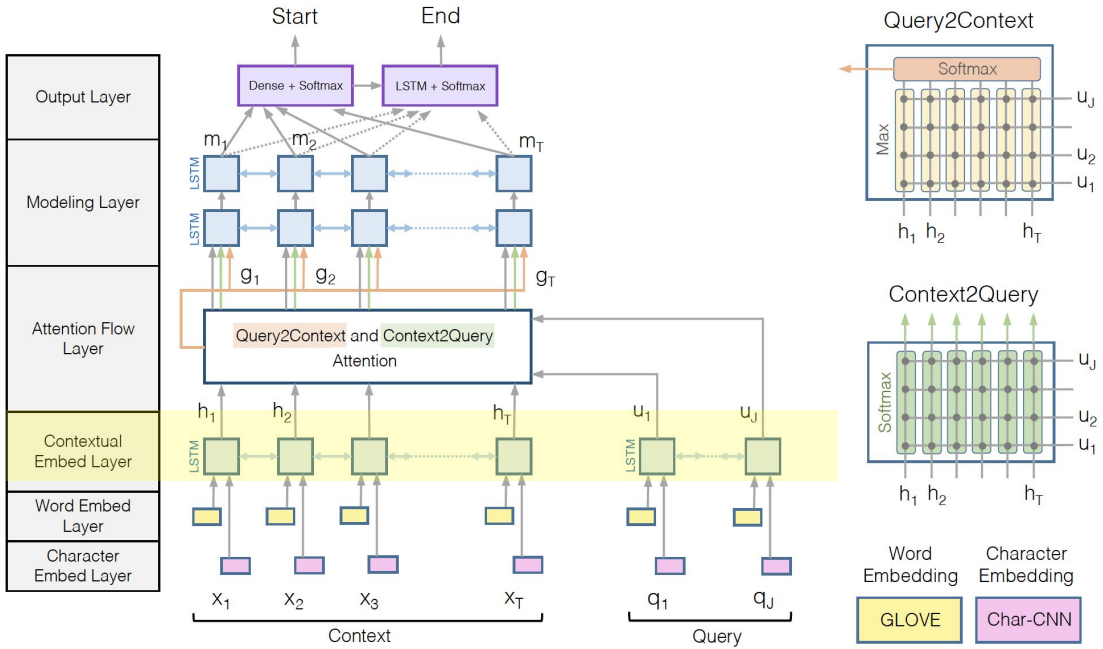
Word Embedding Layer

Maps each word to a vector space using a pre-trained word embedding model.



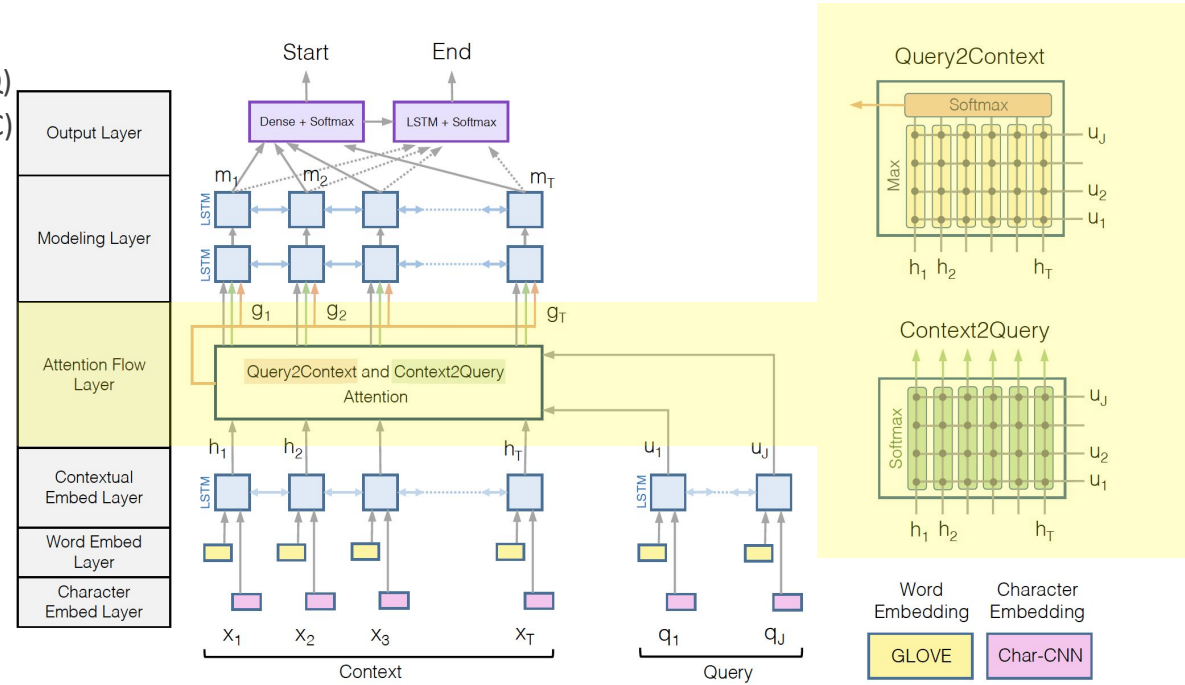
Contextual Embedding Layer

- Utilizes contextual cues from surrounding words to refine the embedding of the words.



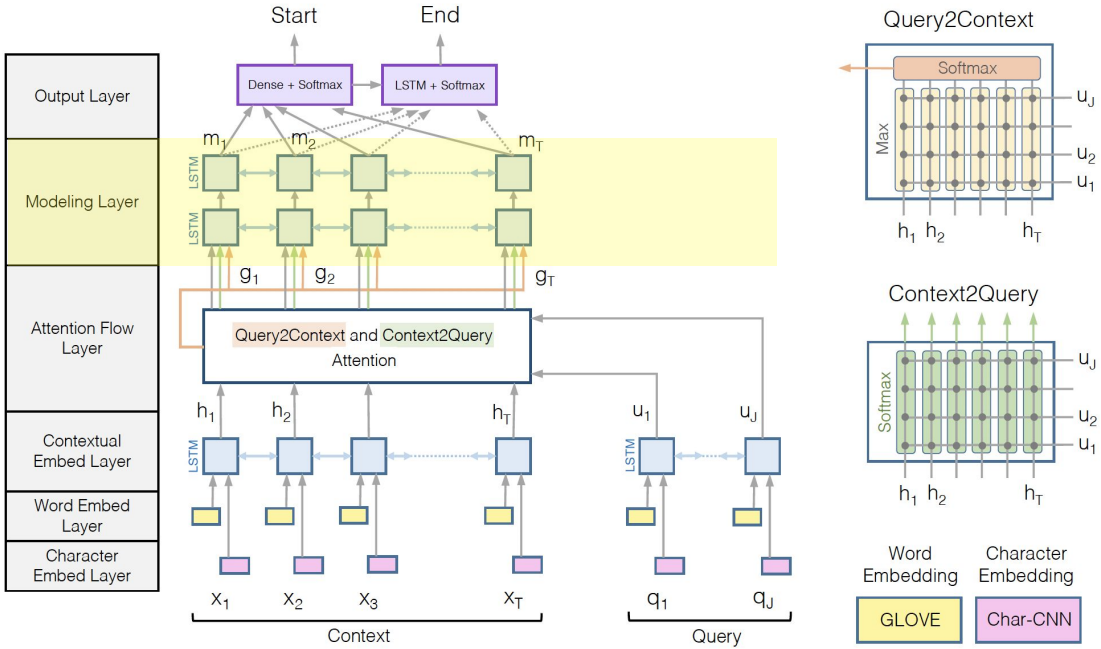
Contextual Embedding Layer

- Similarity matrix is computed:
  - Context-to-query Attention (C2Q)
  - Query-to-context Attention (Q2C)
- Attention vector at each time step flows through to the modeling layer.



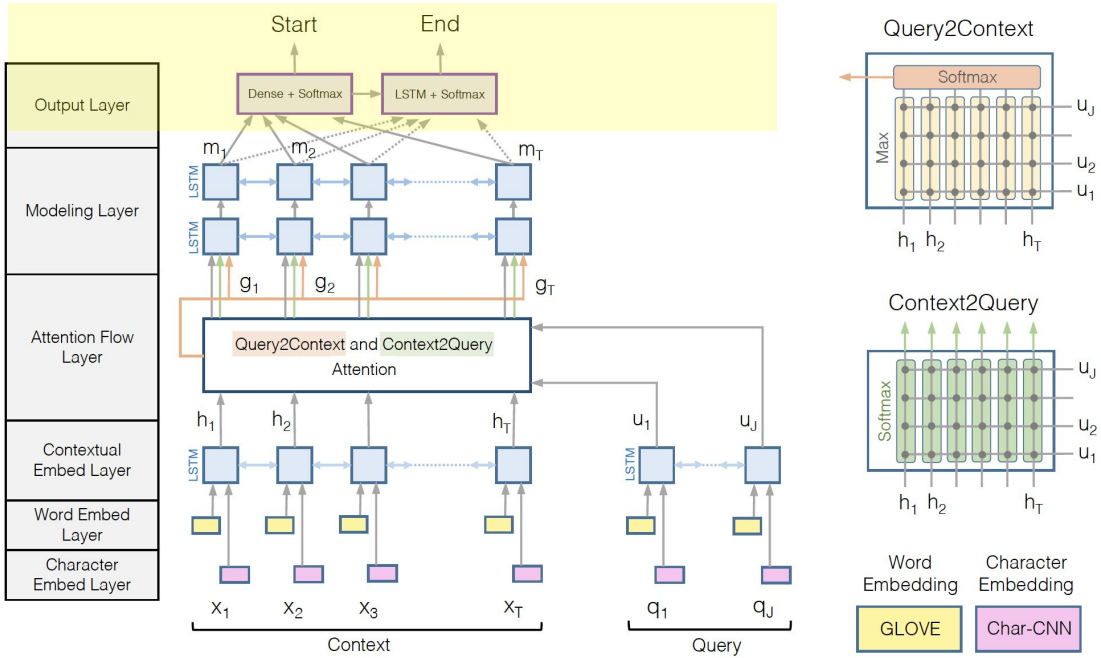
Modeling Layer

- The input to the modeling layer is G, which encodes the query-aware representations of context words.
- The output of the modeling layer captures the interaction among the context words conditioned on the query.
  - This is different from contextual embedding layer.



Output Layer

- The output layer is Application-specific.
- Modular nature of BIDAF allows to easily swap out the output layer based on the task.
- For the QA task, Start and End indices of the answer sub phrase are predicted.



QA Experiment #1

- Dataset:
  - SQuAD
  - Dataset of Wikipedia articles.
  - Answer to each question is always a span in the context
- Metrics:
  - Exact Match (EM)
  - F1 score

https://rajpurkar.github.io/SQuAD-explorer/

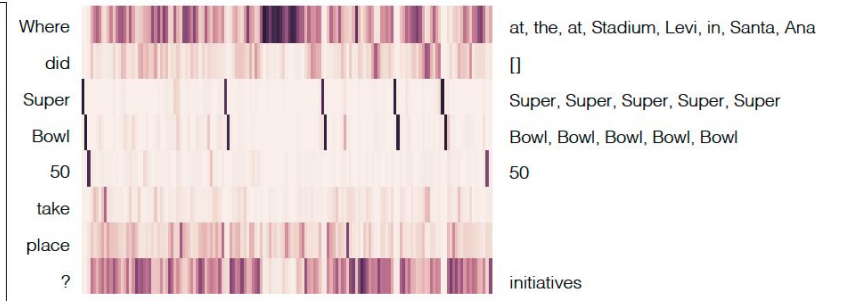
Results

Layer	Query	Closest words in the Context using cosine similarity
Word	When	when, When, After, after, He, he, But, but, before, Before
Contextual	When	When, when, 1945, 1991, 1971, 1967, 1990, 1972, 1965, 1953
Word	Where	Where, where, It, IT, it, they, They, that, That, city
Contextual	Where	where, Where, Rotterdam, area, Nearby, location, outside, Area, across, locations
Word	Who	Who, who, He, he, had, have, she, She, They, they
Contextual	Who	who, whose, whom, Guiscard, person, John, Thomas, families, Elway, Louis
Word	city	City, city, town, Town, Capital, capital, district, cities, province, Downtown
Contextual	city	city, City, Angeles, Paris, Prague, Chicago, Port, Pittsburgh, London, Manhattan
Word	January	July, December, June, October, January, September, February, April, November, March
Contextual	January	January, March, December, August, December, July, July, July, March, December
Word	Seahawks	Seahawks, Broncos, 49ers, Ravens, Chargers, Steelers, quarterback, Vikings, Colts, NFL
Contextual	Seahawks	Seahawks, Broncos, Panthers, Vikings, Packers, Ravens, Patriots, Falcons, Steelers, Chargers
Word	date	date, dates, until, Until, June, July, Year, year, December, deadline
Contextual	date	date, dates, December, July, January, October, June, November, March, February

Table 2: Closest context words to a given query word, using a cosine similarity metric computed in the Word Embedding feature space and the Phrase Embedding feature space.

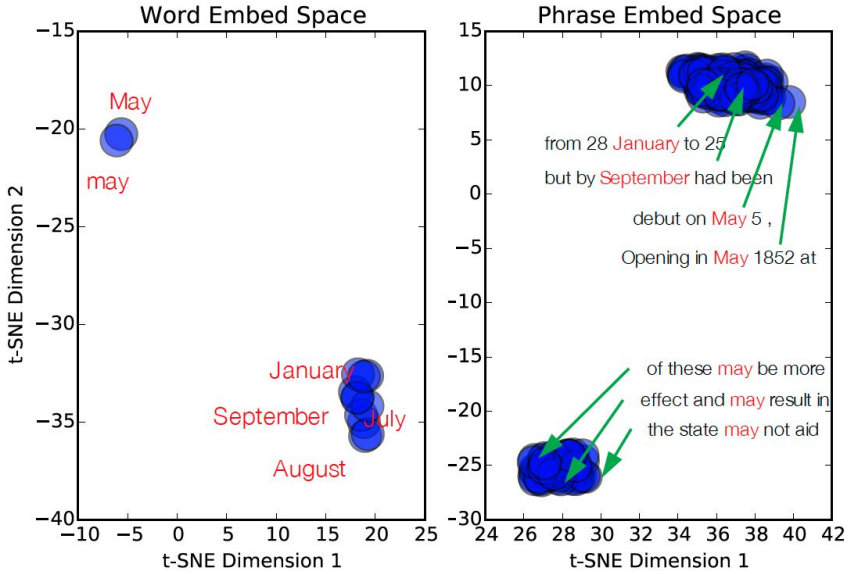
Results

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.



There are 13 natural reserves in Warsaw– among others, Bielany Forest, Kabaty Woods, Czerniaków Lake . About 15 kilometres (9 miles) from Warsaw, the Vistula river's environment changes strikingly and features a perfectly preserved ecosystem, with a habitat of animals that includes the otter, beaver and hundreds of bird species. There are also several lakes in Warsaw – mainly the oxbow lakes, like Czerniaków Lake, the lakes in the Łazienki or Wilanów Parks, Kamionek Lake. There are lot of small lakes in the parks, but only a few are permanent–the majority are emptied before winter to clean them of plants and sediments.

Results


Word Embed Space
May
may
January
September
July
August
t-SNE Dimension 2
t-SNE Dimension 1
Phrase Embed Space
from 28 January to 25
but by September had been
debut on May 5 ,
Opening in May 1852 at
of these may be more
effect and may result in
the state may not aid
t-SNE Dimension 1

Results

	Single Model		Ensemble	
	EM	F1	EM	F1
Logistic Regression Baseline[a]	40.4	51.0	-	-
Dynamic Chunk Reader[b]	62.5	71.0	-	-
Fine-Grained Gating[c]	62.5	73.3	-	-
Match-LSTM[d]	64.7	73.7	67.9	77.0
Multi-Perspective Matching[e]	65.5	75.1	68.2	77.2
Dynamic Coattention Networks[f]	66.2	75.9	71.6	80.4
R-Net[g]	**68.4**	**77.5**	72.1	79.7
BIDAF (Ours)	68.0	77.3	**73.3**	**81.1**

(a) Results on the SQuAD test set

	EM	F1
No char embedding	65.0	75.4
No word embedding	55.5	66.8
No C2Q attention	57.2	67.7
No Q2C attention	63.6	73.7
Dynamic attention	63.5	73.6
BIDAF (single)	67.7	77.3
BIDAF (ensemble)	72.6	80.7

(b) Ablations on the SQuAD dev set

Latest Results

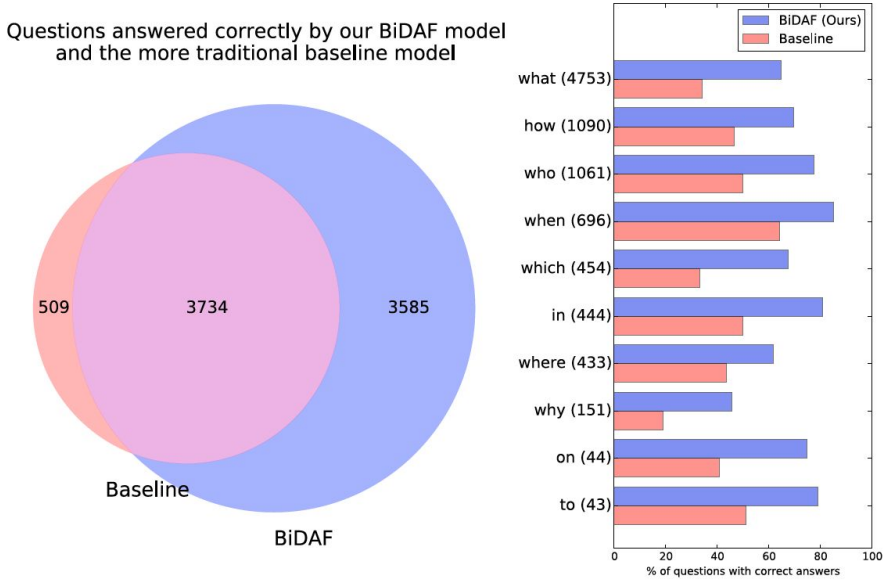
SQuAD1.1 Leaderboard

Here are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Apr 10, 2020	LUKE (single model) *Studio Ousia & NAIST & RIKEN AIP*	**90.202**	**95.379**
2 May 21, 2019	XLNet (single model) *Google Brain & CMU*	89.898	95.080
3 Dec 11, 2019	XLNET-123++ (single model) *MST/EOI* http://tia.today	89.856	94.903
3 Aug 11, 2019	XLNET-123 (single model) *MST/EOI*	89.646	94.930
4 Sep 25, 2019	BERTSP (single model) *NEUKG* http://www.techkg.cn/	88.912	94.584
4 Jul 21, 2019	SpanBERT (single model) *FAIR & UW*	88.839	94.635
5 Jul 03, 2019	BERT+WWM+MT (single model) *Xiaoi Research*	88.650	94.393
18 Nov 17, 2017	BiDAF + Self Attention + ELMo (ensemble) *Allen Institute for Artificial Intelligence*	81.003	87.432
26 Nov 03, 2017	BiDAF + Self Attention + ELMo (single model) *Allen Institute for Artificial Intelligence*	78.580	85.833
20 Feb 13, 2018	BiDAF + Self Attention + ELMo + A2D (single model) *Microsoft Research Asia & NUDT*	79.996	86.711
28 Sep 18, 2018	BiDAF++ with pair2vec (single model) *UW and FAIR*	78.223	85.535
31 Sep 18, 2018	BiDAF++ (single model) *UW and FAIR*	77.573	84.858
39 Feb 13, 2018	SSR-BiDAF *ensemble model*	74.541	82.477
43 Feb 22, 2017	BiDAF (ensemble) *Allen Institute for AI & University of Washington* https://arxiv.org/abs/1611.01603	73.744	81.525

https://rajpurkar.github.io/SQuAD-explorer/

Results


Questions answered correctly by our BiDAF model and the more traditional baseline model
509
3734
3585
Baseline
BiDAF
BiDAF (Ours)
Baseline
what (4753)
how (1090)
who (1061)
when (696)
which (454)
in (444)
where (433)
why (151)
on (44)
to (43)
0
20
40
60
80
100
% of questions with correct answers

CLOZE Text Experiments

- Dataset
  - The reader is asked to fill in word that have been removed from a passage, for measuring one's ability to comprehend text.
  - Each answer in the CNN/DailyMail datasets is always a single word
  - Answer entity might appear more than once in the context paragraph.

- Model
  - All non-entity words in the final classification layer are masked out

Results – Cloze Test

	CNN		DailyMail	
	val	test	val	test
Attentive Reader (Hermann et al., 2015)	61.6	63.0	70.5	69.0
MemNN (Hill et al., 2016)	63.4	6.8	-	-
AS Reader (Kadlec et al., 2016)	68.6	69.5	75.0	73.9
DER Network (Kobayashi et al., 2016)	71.3	72.9	-	-
Iterative Attention (Sordoni et al., 2016)	72.6	73.3	-	-
EpiReader (Trischler et al., 2016)	73.4	74.0	-	-
Stanford AR (Chen et al., 2016)	73.8	73.6	77.6	76.6
GAReader (Dhingra et al., 2016)	73.0	73.8	76.7	75.7
AoA Reader (Cui et al., 2016)	73.1	74.4	-	-
ReasoNet (Shen et al., 2016)	72.9	74.7	77.6	76.6
BIDAF (Ours)	**76.3**	**76.9**	**80.3**	**79.6**
MemNN* (Hill et al., 2016)	66.2	69.4	-	-
ASReader* (Kadlec et al., 2016)	73.9	75.4	78.7	77.7
Iterative Attention* (Sordoni et al., 2016)	74.5	75.7	-	-
GA Reader* (Dhingra et al., 2016)	76.4	77.4	79.1	78.1
Stanford AR* (Chen et al., 2016)	77.2	77.6	80.2	79.2

Related Work - MC

- First Group
  - Dynamic attention mechanism
  - Bahdanau et al.(2015)
    - Attention weights updated dynamically.
  - Hermann et al. (2015)
    - CNN & DailyMail datasets
    - Dynamic attention model performs better than using a single fixed size query vector.
  - Wang & Jiang (2016)
    - Reverse the direction of attention

- In contrast, BIDAF uses a memory-less attention mechanism.

Related Work - MC

- Second Group
  - Attention weights are computed once
    - Kadlec et al. (2016)
  - Cui et al.(2016)
    - 2D similarity matrix for computing query-to-context attention.

- BIDAF
  - Lets the attention vectors flow into the modeling layer.

Related Work - MC

- Third Group
  - Multi-hop:
    - Repeats computing an attention vector between the query and the context through multiple layers
    - Sordoni et al., 2016; Dhingra et al., 2016)
  - Shen et al. (2016)
    - Combines Memory networks with RL to dynamically control the number of hops.

- BIDAF
  - Can be extended to incorporate multiple hops.

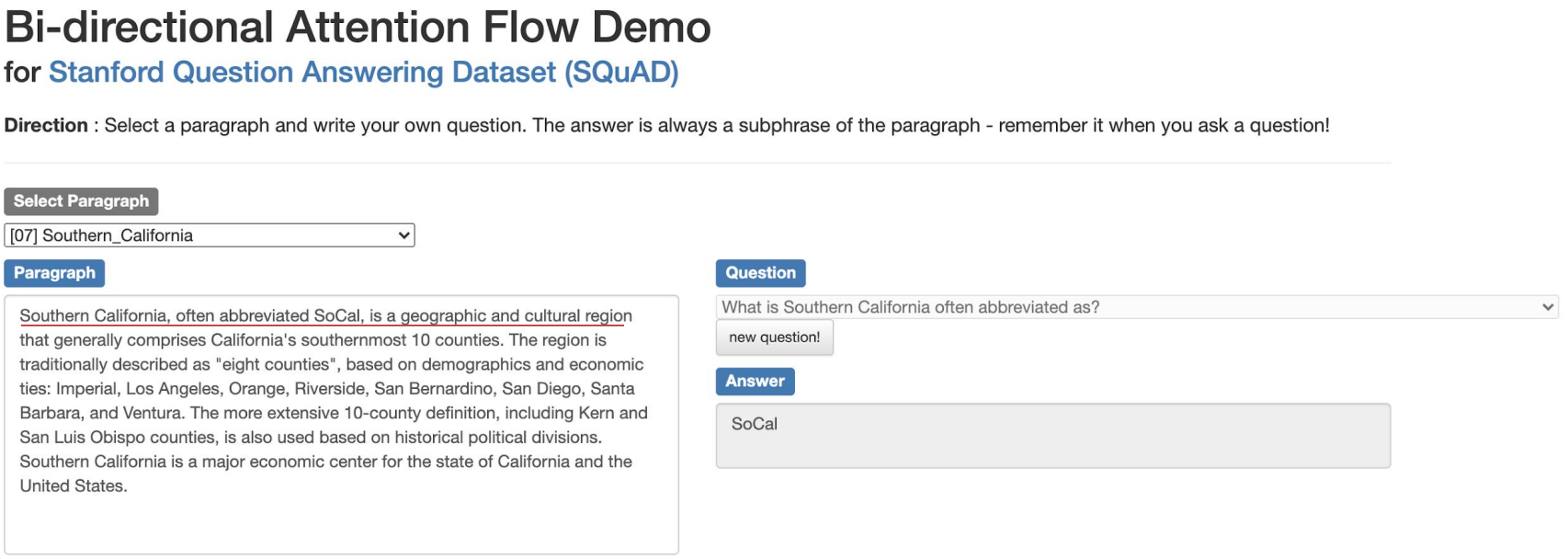
Related Work - VQA

- Coarse Level
  - Zhu et al.; Xiong et al. (2016)
  - Question attends to different patches of image.
- Finer level
  - Each question word attends to each image patch
  - Xu & Saenko (2016)
- In addition to attending from question to image patches, attend from the image back to the question words.
  - Lu et al.(2016)

Discussion

- With the advent of transformers such as BERT, How do we position BIDAF?
- BIDAF uses significantly less parameters than BERT, Hence BIDAF based approaches can be used when we have constraints on the total number of parameters/ computation cost.
- Does the pluggable final output layer make BIDAF better for transfer learning?

DEMO



https://allenai.github.io/bi-att-flow/

THANK YOU

NOVEMBER 2020

NEURAL MODULE NETWORKS FOR REASONING OVER TEXT

Nitish Gupta, Kevin Lin , Dan Roth, Sameer Singh & Matt Gardner at **ICLR 2020**

VIKRAM MANDIKAL
Department of Computer Science
The University of Texas at Austin

Outline

- Motivation
- Overview
- Neural Module Network
- Modules
- Auxiliary Loss and Intermediate Supervision
- Results

Outline

- **Motivation**
- Overview
- Neural Module Network
- Modules
- Auxiliary Loss and Intermediate Supervision
- Results

Motivation

DROP dataset

Reasoning	**Passage** (some parts shortened)	**Question**	**Answer**	**BiDAF**
Subtraction (28.8%)	That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000	$16.3 million
Comparison (18.2%)	In **1517, the seventeen-year-old King sailed to Castile**. There, his Flemish court .... **In May 1518, Charles traveled to Barcelona in Aragon**.	Where did Charles travel to first, Castile or Barcelona?	Castile	Aragon
Selection (19.4%)	In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, **Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack** to tell the story of the events that led up to the battle.	Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?	Don Mueller	Baker

Motivation

DROP dataset

Addition (11.7%)	Before the UNPROFOR fully deployed, the HV clashed with an armed force of the RSK in the village of Nos Kalik, located in a pink zone near Šibenik, and captured the village at 4:45 p.m. on **2 March 1992**. The JNA formed a battlegroup to counterattack the **next day**.	What date did the JNA form a battlegroup to counterattack after the village of Nos Kalik was captured?	3 March 1992	2 March 1992
Count (16.5%) and Sort (11.7%)	Denver would retake the lead with kicker **Matt Prater nailing a 43-yard field goal**, yet Carolina answered as kicker **John Kasay ties the game with a 39-yard field goal**. … Carolina closed out the half with **Kasay nailing a 44-yard field goal**. … In the fourth quarter, Carolina sealed the win with **Kasay's 42-yard field goal**.	Which kicker kicked the most field goals?	John Kasay	Matt Prater
Coreference Resolution (3.7%)	**James Douglas** was the second son of Sir George Douglas of Pittendreich, and Elizabeth Douglas, daughter David Douglas of Pittendreich. Before **1543 he married Elizabeth**, daughter of James Douglas, 3rd Earl of Morton. **In 1553 James Douglas succeeded to the title and estates of his father-in-law**.	How many years after he married Elizabeth did James Douglas succeed to the title and estates of his father-in-law?	10	1553

Motivation

DROP dataset

Other Arithmetic (3.2%)	Although the movement initially gathered some **60,000 adherents**, the subsequent establishment of the Bulgarian Exarchate **reduced their number by some 75%**.	How many adherents were left after the establishment of the Bulgarian Exarchate?	15000	60,000
Set of spans (6.0%)	According to some sources 363 civilians were killed in **Kavadarci**, 230 in **Negotino** and 40 in **Vatasha**.	What were the 3 villages that people were killed in?	Kavadarci, Negotino, Vatasha	Negotino and 40 in Vatasha
Other (6.8%)	This **Annual Financial Report** is our principal financial statement of accountability. The **AFR gives a comprehensive view** of the Department's financial activities ...	What does AFR stand for?	Annual Financial Report	one of the Big Four audit firms

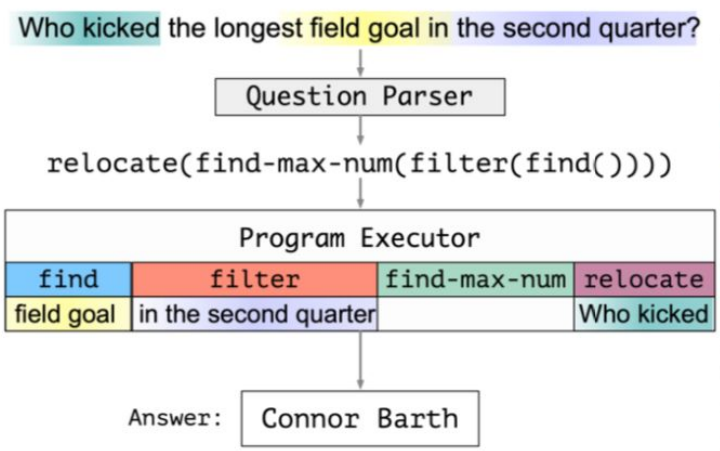
Motivation

- Neural network modules are interpretable, modular and inherently compositional in nature - used in VQA domains.
- Extend Neural Module Networks for question answering against open-domian text.
- Introduce **probabilistic and differentiable modules** to reason over text.
- Can supervise the intermediate latent decisions in compositional question by using a proposed auxiliary loss.

Outline

- Motivation
- **Overview**
- Neural Module Network
- Modules
- Auxiliary Loss and Intermediate Supervision
- Results

Overview



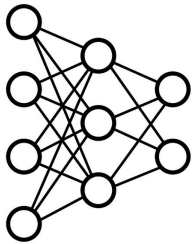
In the first quarter, Buffalo trailed as Chiefs QB Tyler Thigpen completed a 36-yard TD pass to RB Jamaal Charles. The Bills responded with RB Marshawn Lynch getting a 1-yard touchdown run. In the second quarter, Buffalo took the lead as kicker Rian Lindell made a 21-yard and a 40-yard field goal. Kansas City answered with Thigpen completing a 2-yard TD pass. Buffalo regained the lead as Lindell got a 39-yard field goal. The Chiefs struck with kicker Connor Barth getting a 45-yard field goal, yet the Bills continued their offensive explosion as Lindell got a 34-yard field goal, along with QB Edwards getting a 15-yard TD run. In the third quarter, Buffalo continued its poundings with Edwards getting a 5-yard TD run, while Lindell got himself a 48-yard field goal. Kansas City tried to rally as Thigpen completed a 45-yard TD pass to WR Mark Bradley, yet the Bills replied with Edwards completing an 8-yard TD pass to WR Josh Reed. In the fourth quarter, Edwards completed a 17-yard TD pass to TE Derek Schouman.

Outline

- Motivation
- Overview
- **Neural Module Network**
- Modules
- Auxiliary Loss and Intermediate Supervision
- Results

Neural Module Networks (NMNs)

What is this?

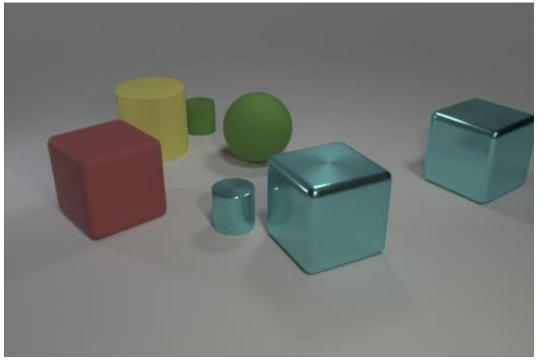


Neural network

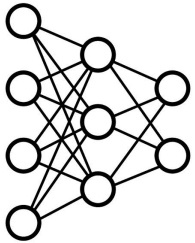
Cat

Image recognition problem can be solved by a single large network which can map (image, question) to answer.

Neural Module Networks (NMNs)



What is the color of thing with the same size as the blue cylinder?

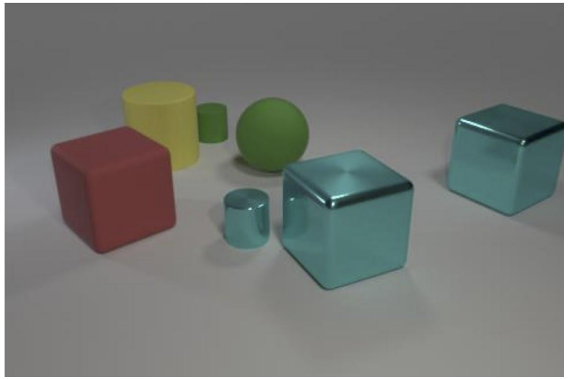


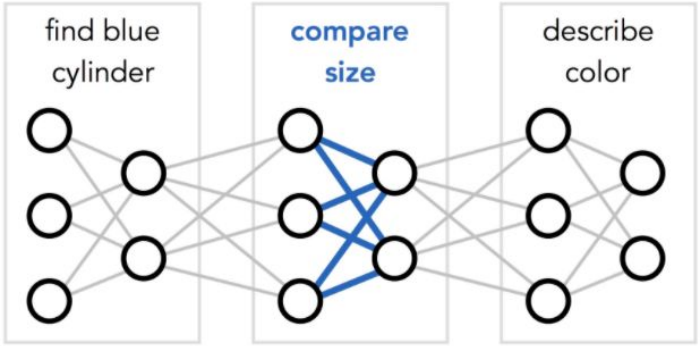
Neural network

X

This requires *reasoning*, cannot be performed by a fixed architecture network.

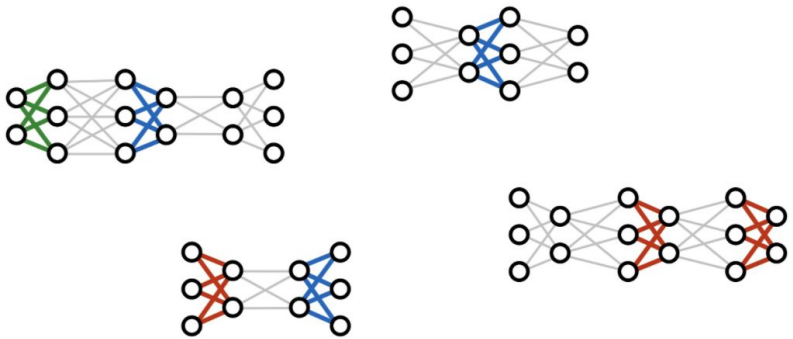
Neural Module Networks (NMNs)





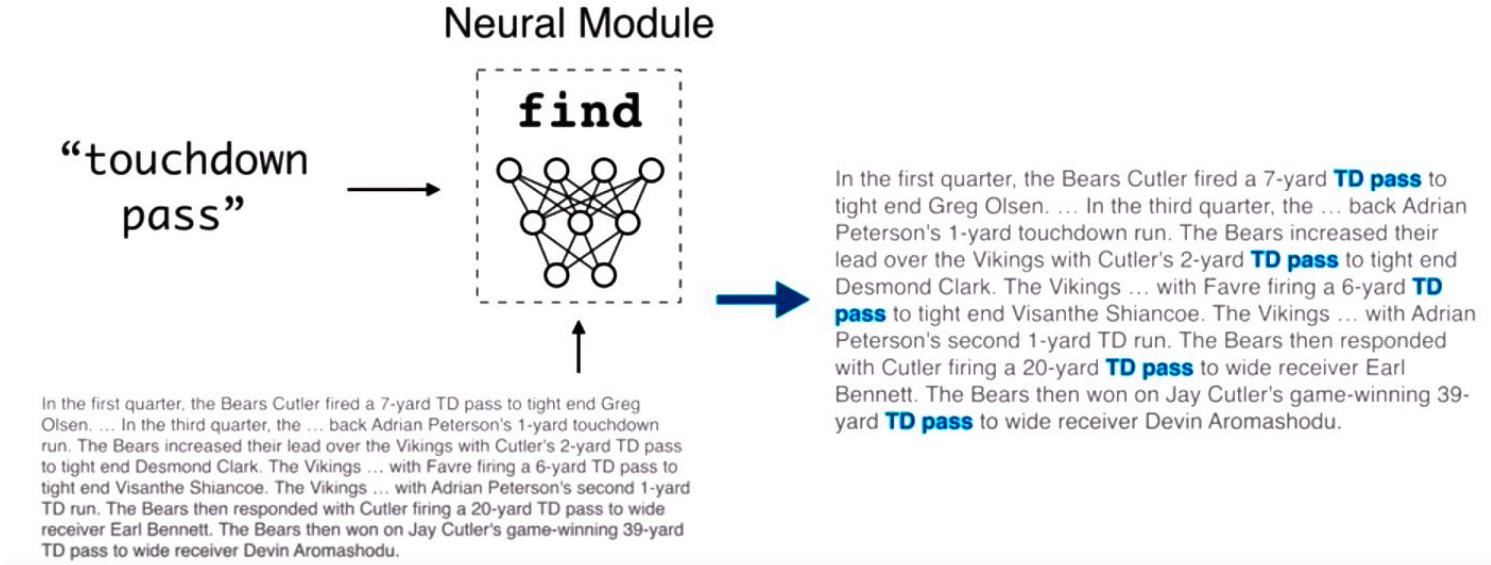
green

Neural Module Networks (NMNs)



- A collection of **“neural modules”** instead of a single large neural network.
- Each neural module implements a single step of reasoning.
- Neural modules are assembled dynamically according to the question.
- Best of both worlds: the flexibility and interpretability of discrete compositionality, combined with the representational power of deep networks

Neural Module Networks (NMNs)



Learns how to execute this function

Neural Module Networks (NMNs)

- **Modules:** Define various modules such as find, filter etc for different data types.
- **Contextual Token representations:** obtained through bi-directional GRU or pre-trained BERT.
- **Question parser:** Encoder-decoder model to map the question into an executable program.

Outline

- Motivation
- Overview
- Neural Module Network
- **Modules**
- Auxiliary Loss and Intermediate Supervision
- Results

Modules

- Perform various natural language tasks and symbolic reasoning tasks.
- Designed to work in a **probabilistic and differentiable manner.**

	Module	In	Out	Task
Natural language reasoning	find	Q	P	For question spans in the input, find similar spans in the passage
	filter	Q, P	P	Based on the question, select a subset of spans from the input
	relocate	Q, P	P	Find the argument asked for in the question for input paragraph spans
	find-num	P	N	} Find the number(s) / date(s) associated to the input paragraph spans
	find-date	P	D	
Symbolic reasoning	count	P	C	Count the number of input passage spans
	compare-num-lt	P, P	P	Output the span associated with the smaller number.
	time-diff	P, P	TD	Difference between the dates associated with the paragraph spans
	find-max-num	P	P	Select the span that is associated with the largest number
	span	P	S	Identify a contiguous span from the attended tokens

Modules

`find(Q) → P`

Input: Distribution over question tokens
Output: Distribution over passage tokens

`find("How many field goals?")`

Denver would retake the lead with kicker Matt Prater nailing a 43-yard field goal, yet Carolina answered as kicker John Kasay ties the game with a 39-yard field goal… Carolina closed out the half with Kasay nailing a 44-yard field goal… In the fourth quarter, Carolina sealed the win with Kasay's 42-yard field goal.

Modules

`find-num(P) → N`

Input: Distribution over passage tokens
Output: Distribution over passage numbers

`find- num(`

Denver would retake the lead with kicker Matt Prater nailing a 43-yard field goal, yet Carolina answered as kicker John Kasay ties the game with a 39-yard field goal… Carolina closed out the half with Kasay nailing a 44-yard field goal… In the fourth quarter, Carolina sealed the win with Kasay's 42-yard field goal.

`)`

Denver would retake the lead with kicker Matt Prater nailing a 43-yard field goal, yet Carolina answered as kicker John Kasay ties the game with a 39-yard field goal… Carolina closed out the half with Kasay nailing a 44-yard field goal… In the fourth quarter, Carolina sealed the win with Kasay's 42-yard field goal.

Modules

$\max(N) \rightarrow N$

Input: Distribution over passage numbers
Output: Distribution over passage numbers

max(22, 9, 3, 62, 15, 3)

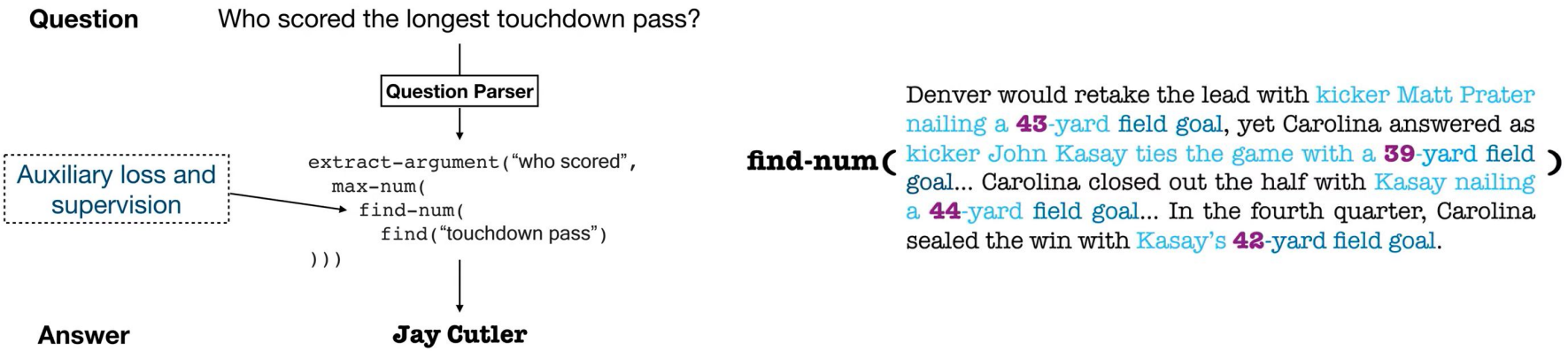
22, 9, 3, 62, 15, 3

Parameter-free module with a differentiable & analytical formulation

Outline

- Motivation
- Overview
- Neural Module Network
- Modules
- **Auxiliary Loss and Intermediate Supervision**
- Results

Auxiliary Loss and Intermediate Supervision



Auxiliary Loss and Intermediate Supervision

- **Unsupervised auxiliary loss:** To induce that that the arguments of a mention should appear near it (*in find-num, fi*

$$H^{\text{n}}_{\text{loss}} = -\sum_{i=1}^{m} \log\left(\sum_{j=0}^{N_{\text{tokens}}} \mathbb{1}_{n_j \in [i \pm W]} \mathbf{A}^{\text{num}}{}_{ij}\right)$$

find-num(Denver would retake the lead with kicker Matt Prater nailing a 43-yard field goal, yet Carolina answered as kicker John Kasay ties the game with a 39-yard field goal... Carolina closed out the half with Kasay nailing a **44**-yard **field goal**. In the fourth quarter, Carolina sealed the win with Kasay's 42-yard field goal. **)**

Outline

- Motivation
- Overview
- Neural Module Network
- Modules
- Auxiliary Loss and Intermediate Supervision
- **Results**

Results

- **DATASET:**
  - Discrete Reasoning over Paragraphs(DROP)
  - 20,000 training/validation and 1800 testing.
  - Different types of questions:
    - **Date-Compare** e.g. What happened last, commission being granted to Robert or death of his cousin?
    - **Date-Difference** e.g. How many years after his attempted assassination was James II coronated?
    - **Number-Compare** e.g. Were there more of cultivators or main agricultural labourers in Sweden?
    - **Extract-Number** e.g. How many yards was Kasay's shortest field goal during the second half?
    - **Count** e.g. How many touchdowns did the Vikings score in the first half?
    - **Extract-Argument** e.g. Who threw the longest touchdown pass in the first quarter?

Results

Model	F1	EM
NAQANET	62.1	57.9
TAG-NABERT+	74.2	70.6
NABERT+	75.4	72.0
MTMSN	76.5	73.1
OUR MODEL (W/ GRU)	73.1	69.6
OUR MODEL (W/ BERT)	**77.4**	**74.0**

(a) **Performance on DROP (pruned)**

Question Type	MTMSN	Our Model (w/ BERT)
DATE-COMPARE (18.6%)	**85.2**	82.6
DATE-DIFFERENCE (17.9%)	72.5	**75.4**
NUMBER-COMPARE (19.3%)	85.1	**92.7**
EXTRACT-NUMBER (13.5%)	80.7	**86.1**
COUNT (17.6%)	**61.6**	55.7
EXTRACT-ARGUMENT (12.8%)	66.6	**69.7**

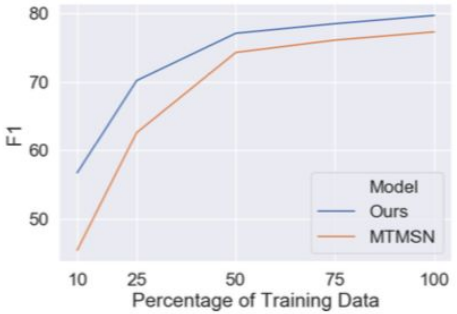
(b) **Performance by Question Type (F1)**

Table 2: **Performance of different models on the dataset and across different question types.**

Results

Supervision Type		w/ BERT	w/ GRU
H_{loss}	MOD-SUP		
✓	✓	**77.4**	**73.1**
✓		76.3	71.8
	✓	–*	57.3

(a) **Effect of Auxiliary Supervision:** The auxiliary loss contributes significantly to the performance, whereas module output supervision has little effect. **Training diverges without H_{loss} for the BERT-based model.*

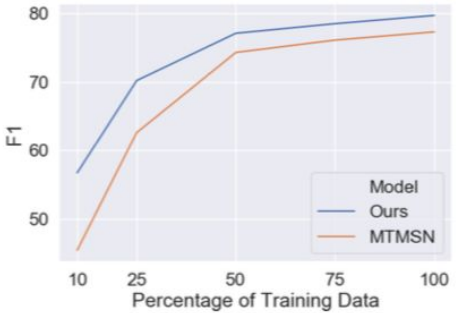


(b) **Performance with less training data**: Our model performs significantly better than the baseline with less training data, showing the efficacy of explicitly modeling compositionality.

Results

Supervision Type		w/ BERT	w/ GRU
H_{loss}	MOD-SUP		
✓	✓	**77.4**	**73.1**
✓		76.3	71.8
	✓	–*	57.3

(a) **Effect of Auxiliary Supervision:** The auxiliary loss contributes significantly to the performance, whereas module output supervision has little effect. **Training diverges without H_{loss} for the BERT-based model.*



(b) **Performance with less training data**: Our model performs significantly better than the baseline with less training data, showing the efficacy of explicitly modeling compositionality.

Related Work

- Multi-Type Multi-Span Network (**MTMSN**) - a neural reading comprehension model that combines a multi-type answer predictor designed to support various answer types (e.g., span, count, negation, and arithmetic expression).

- **NAQANet**: produces three answer types: (1) spans from the question; (2) counts; (3) addition or subtraction over numbers.

Discussion

- How to generalize this approach for more diverse reasoning tasks? Is this a scalable approach?
- Are these modules tailored for the DROP dataset?
- The individual performance of each of the modules should be discussed - maybe on a subset of the test-set.

Discussion

- How do we know that the modules are performing the intended task - results on this will support the interpretability claim.
- It would be useful to analyze what proportion of the failures are due to the parser and the modules.
- Is it fair to compare with MTMSM on a selected subset of DROP, as MTMSM is designed to handle a broader set of questions? (also this approach uses additional supervision signals!)

Future Directions

- Context conditional parsing: Currently the parsing cannot handle context-conditional parsing.
- Structured parsing is restrictive and cannot fully capture the diverse semantics in natural language.
- As we have seen, this approach is better at handling selected questions, hence can we combine it with MTMSM to handle a wider variety of questions better?

References

•Gupta, Nitish, et al. "Neural module networks for reasoning over text." *arXiv preprint arXiv:1912.04971* (2019).

•A talk: https://www.youtube.com/watch?v=_yp4VhXV_2g&t=2124s

NOVEMBER 2020

TEXAS
The University of Texas at Austin

THANK YOU!