Did the Model Understand the Question?

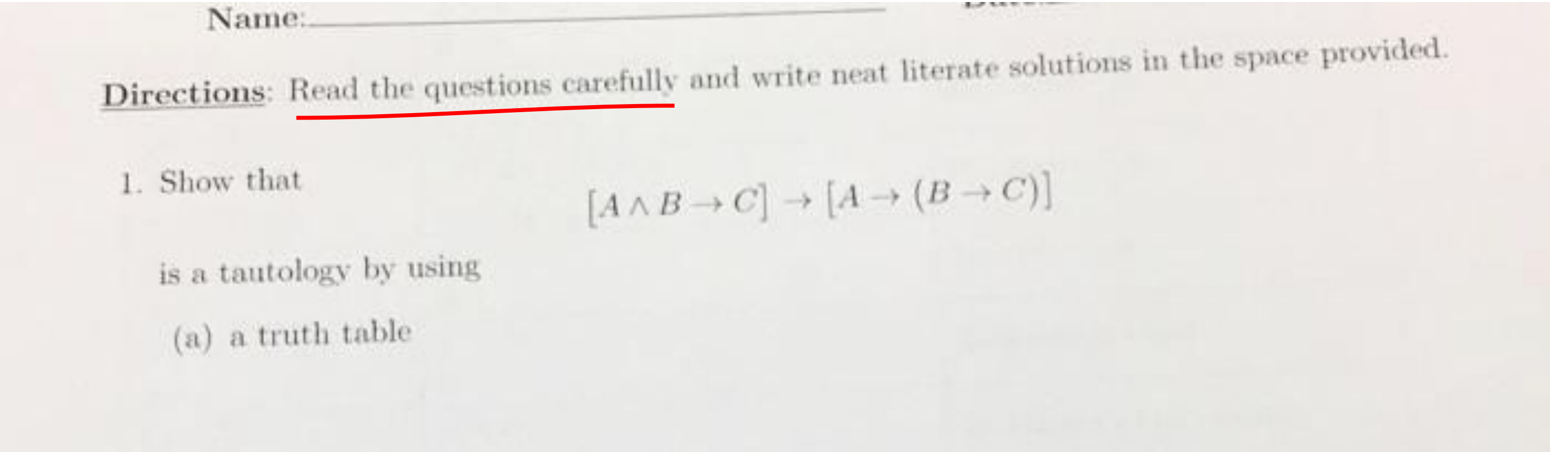
Pramod Kaushik Mudrakarta

THE UNIVERSITY OF CHICAGO
PhD Student

Google
Intern

joint work with Ankur Taly (G🧠), Mukund Sundararajan (G), and Kedar Dhamdhere (G)

Read the question carefully!



Name:____________________

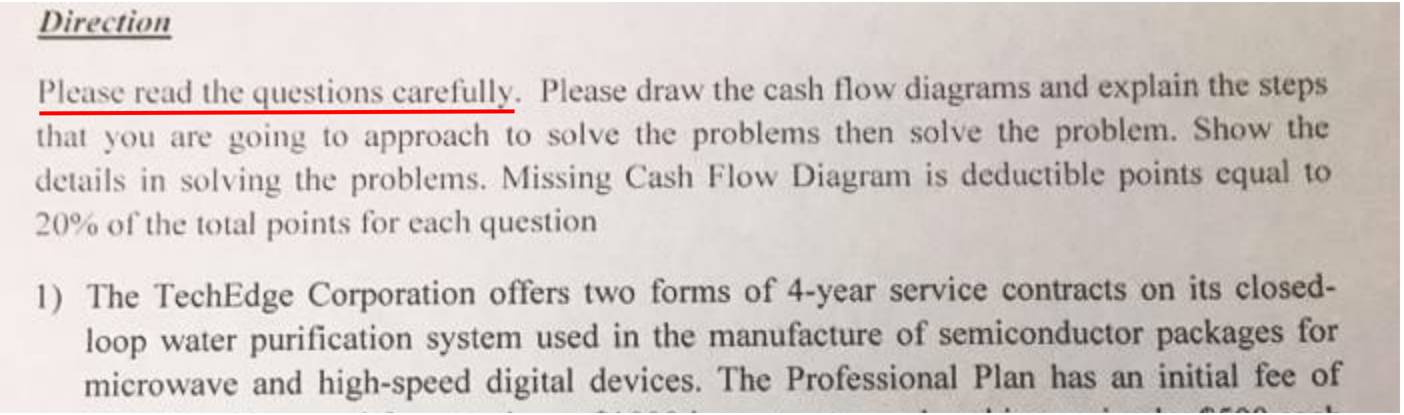
Directions: Read the questions carefully and write neat literate solutions in the space provided.

1. Show that

$$[A \wedge B \to C] \to [A \to (B \to C)]$$

is a tautology by using

(a) a truth table

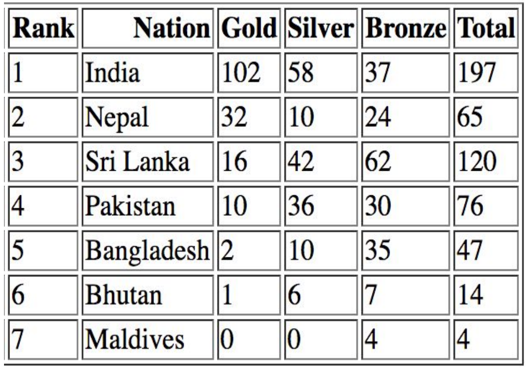


Direction

Please read the questions carefully. Please draw the cash flow diagrams and explain the steps that you are going to approach to solve the problems then solve the problem. Show the details in solving the problems. Missing Cash Flow Diagram is deductible points equal to 20% of the total points for each question

1) The TechEdge Corporation offers two forms of 4-year service contracts on its closed-loop water purification system used in the manufacture of semiconductor packages for microwave and high-speed digital devices. The Professional Plan has an initial fee of

Tabular QA

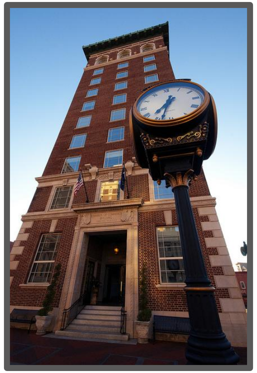


Rank	Nation	Gold	Silver	Bronze	Total
1	India	102	58	37	197
2	Nepal	32	10	24	65
3	Sri Lanka	16	42	62	120
4	Pakistan	10	36	30	76
5	Bangladesh	2	10	35	47
6	Bhutan	1	6	7	14
7	Maldives	0	0	4	4

Q: How many medals did India win?
A: 197

Neural Programmer (2016)
33.5% accuracy on
WikiTableQuestions (state of the art)

Visual QA



Q: How symmetrical are the white bricks on either side of the building?
A: very

Kazemi and Elqursh (2017) model.
61.1% on VQA 1.0 dataset
(state of the art = 66.7%)

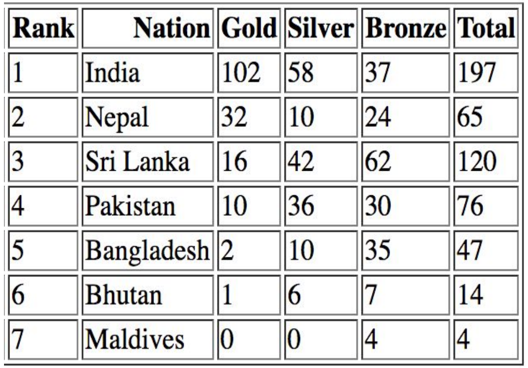
Reading Comprehension

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager

Q: What is the name of the quarterback who was 38 in Super Bowl XXXIII?
A: John Elway

Yu et al (2018) model.
84.6 F-1 score on SQuAD
(state of the art)

Tabular QA

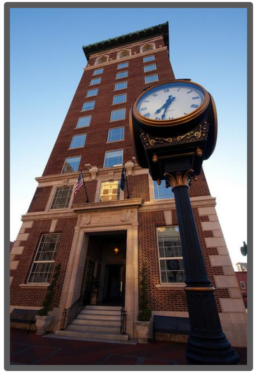


Rank	Nation	Gold	Silver	Bronze	Total
1	India	102	58	37	197
2	Nepal	32	10	24	65
3	Sri Lanka	16	42	62	120
4	Pakistan	10	36	30	76
5	Bangladesh	2	10	35	47
6	Bhutan	1	6	7	14
7	Maldives	0	0	4	4

Q: How many medals did India win?
A: 197

Neural Programmer (2016)
33.5% accuracy on
WikiTableQuestions (state of the art)

Visual QA



Q: How symmetrical are the white bricks on either side of the building?
A: very

Kazemi and Elqursh (2017) model.
61.1% on VQA 1.0 dataset
(state of the art = 66.7%)

Reading Comprehension

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager

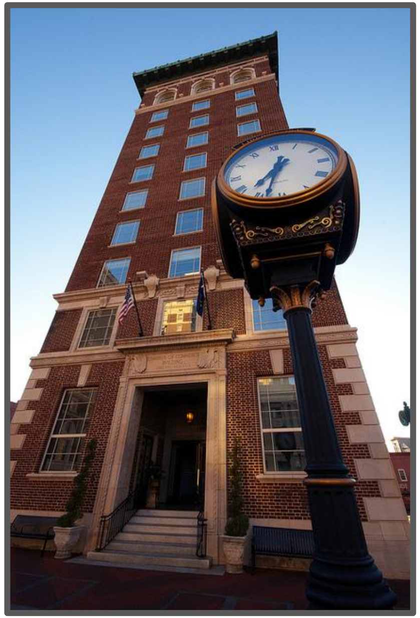
Q: What is the name of the quarterback who was 38 in Super Bowl XXXIII?
A: John Elway

Yu et al (2018) model.
84.6 F-1 score on SQuAD
(state of the art)

Have the models read the question carefully?

Visual QA

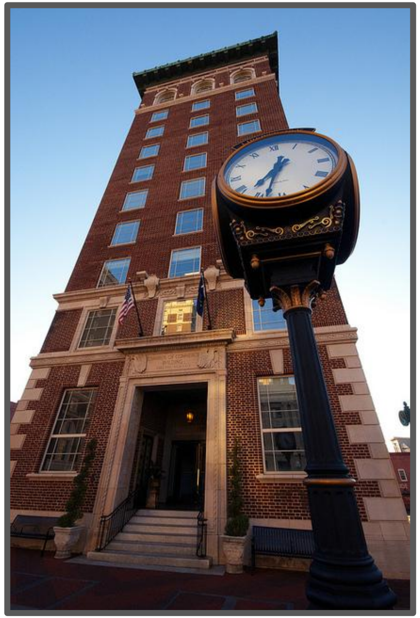
Kazemi and Elqursh (2017) model.
61.1% on VQA dataset (state of the art = 66.7%)



Q: How symmetrical are the white bricks on either side of the building?
A: very

Visual QA

Kazemi and Elqursh (2017) model.
61.1% on VQA dataset (state of the art = 66.7%)

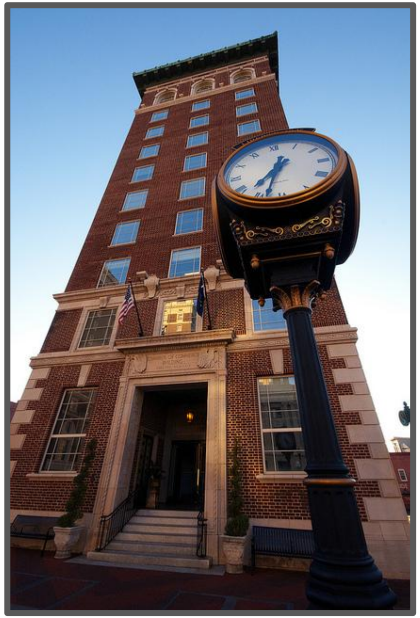


Q: How symmetrical are the white bricks on either side of the building?
A: very

Q: How asymmetrical are the white bricks on either side of the building?
A: very

Visual QA

Kazemi and Elqursh (2017) model.
61.1% on VQA dataset (state of the art = 66.7%)



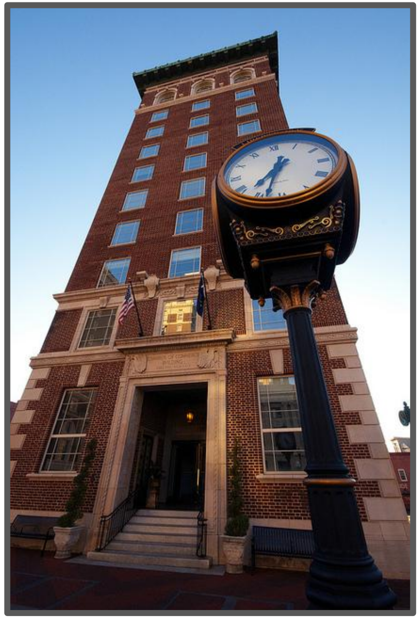
Q: How symmetrical are the white bricks on either side of the building?
A: very

Q: How asymmetrical are the white bricks on either side of the building?
A: very

Q: How big are the white bricks on either side of the building?
A: very

Visual QA

Kazemi and Elqursh (2017) model.
61.1% on VQA dataset (state of the art = 66.7%)



Q: How symmetrical are the white bricks on either side of the building?
A: very

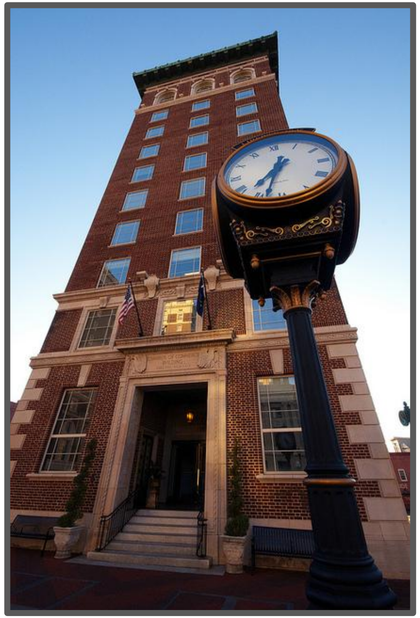
Q: How asymmetrical are the white bricks on either side of the building?
A: very

Q: How big are the white bricks on either side of the building?
A: very

Q: How spherical are the white bricks on either side of the building?
A: very

Visual QA

Kazemi and Elqursh (2017) model.
61.1% on VQA dataset (state of the art = 66.7%)



Q: How symmetrical are the white bricks on either side of the building?
A: very

Q: How asymmetrical are the white bricks on either side of the building?
A: very

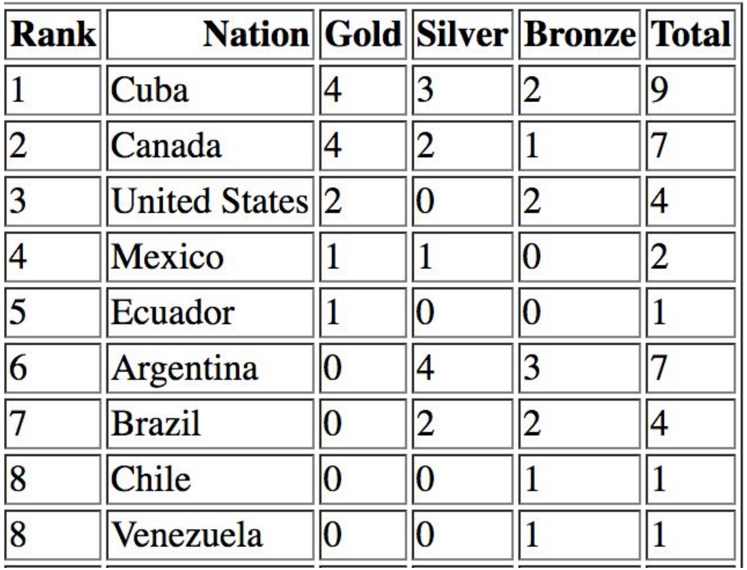
Q: How big are the white bricks on either side of the building?
A: very

Q: How spherical are the white bricks on either side of the building?
A: very

Q: How fast are the bricks speaking on either side of the building?
A: very

QA over tables

Neural Programmer (2016)
33.5% validation accuracy on WikiTableQuestions dataset (state of the art)



Rank	Nation	Gold	Silver	Bronze	Total
1	Cuba	4	3	2	9
2	Canada	4	2	1	7
3	United States	2	0	2	4
4	Mexico	1	1	0	2
5	Ecuador	1	0	0	1
6	Argentina	0	4	3	7
7	Brazil	0	2	2	4
8	Chile	0	0	1	1
8	Venezuela	0	0	1	1

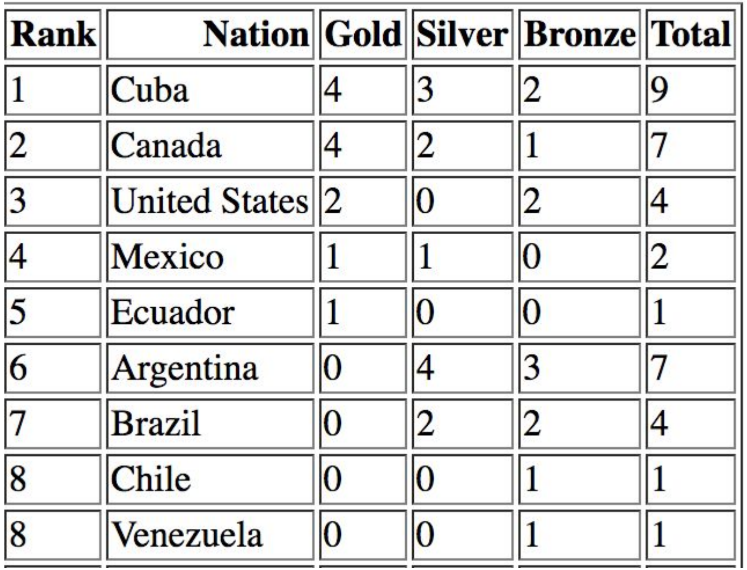
Q: Which country won the most medals?

Neural Programmer:

```
max(total), print(nation)
```

QA over tables

Neural Programmer (2016)
33.5% validation accuracy on WikiTableQuestions dataset (state of the art)



Rank	Nation	Gold	Silver	Bronze	Total
1	Cuba	4	3	2	9
2	Canada	4	2	1	7
3	United States	2	0	2	4
4	Mexico	1	1	0	2
5	Ecuador	1	0	0	1
6	Argentina	0	4	3	7
7	Brazil	0	2	2	4
8	Chile	0	0	1	1
8	Venezuela	0	0	1	1

Q: Which country won the most medals?

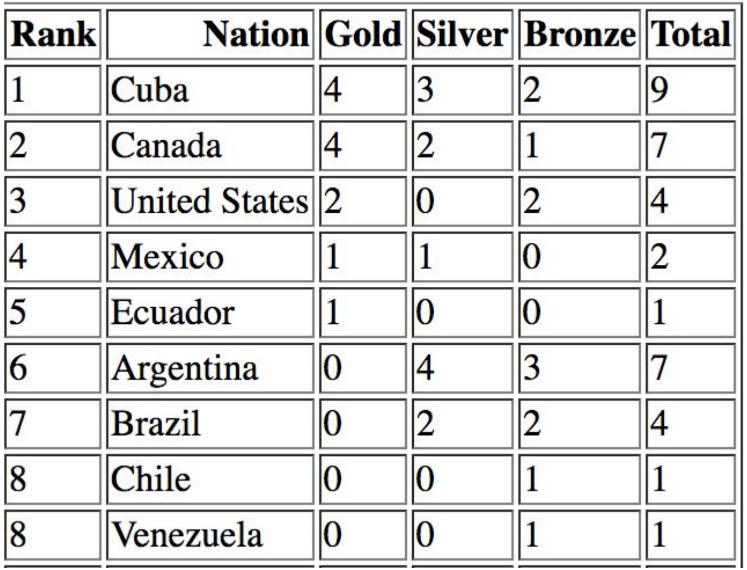
Neural Programmer:

```
max(total), print(nation)
```

A: Cuba ✓

QA over tables

Neural Programmer (2016)
33.5% validation accuracy on WikiTableQuestions dataset (state of the art)



Rank	Nation	Gold	Silver	Bronze	Total
1	Cuba	4	3	2	9
2	Canada	4	2	1	7
3	United States	2	0	2	4
4	Mexico	1	1	0	2
5	Ecuador	1	0	0	1
6	Argentina	0	4	3	7
7	Brazil	0	2	2	4
8	Chile	0	0	1	1
8	Venezuela	0	0	1	1

Q: Which country won the most number of medals?

Neural Programmer:

```
max(bronze), print(nation)
```

A: Argentina ✗

Test/dev accuracy does not show us the entire picture

Jia and Liang (2017): Adversarial Attacks on Reading Comprehension Models

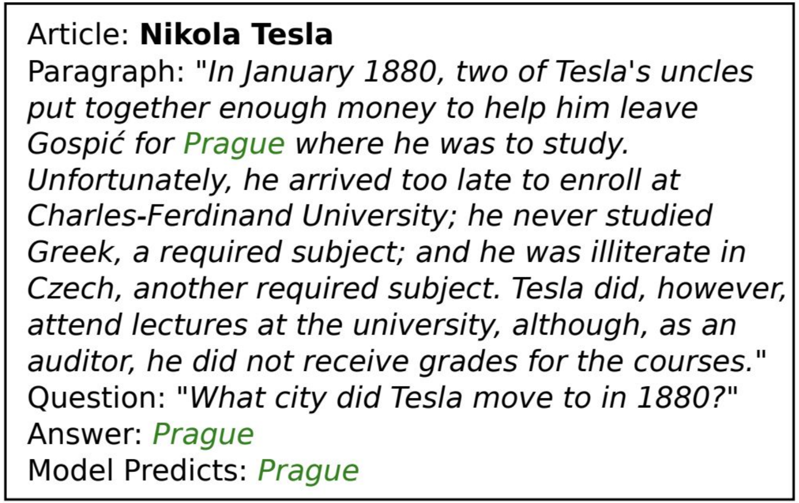
EMNLP 2017 Outstanding Paper Award

Add an adversarial sentence to the paragraph to fool the model

Jia and Liang (2017): Adversarial Attacks on Reading Comprehension Models

EMNLP 2017 Outstanding Paper Award

Add an adversarial sentence to the paragraph to fool the model



Article: **Nikola Tesla**

Paragraph: "*In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.*"

Question: "*What city did Tesla move to in 1880?*"

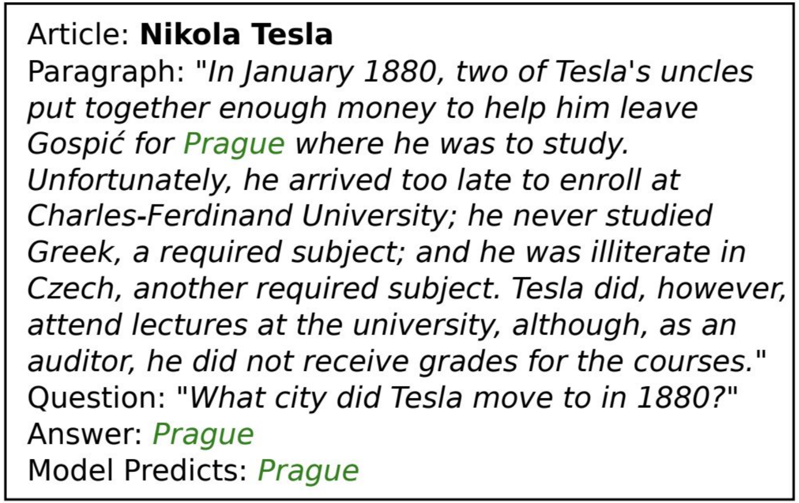
Answer: *Prague*

Model Predicts: *Prague*

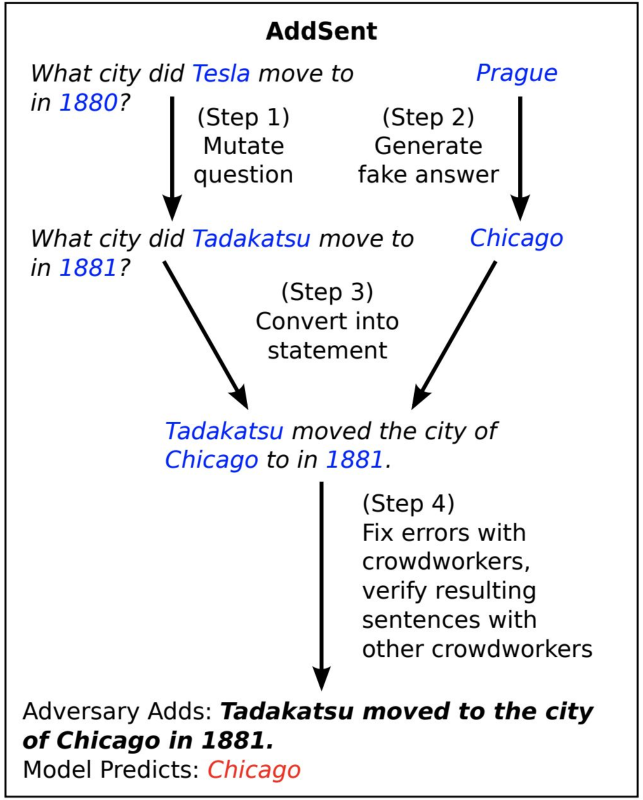
Jia and Liang (2017): Adversarial Attacks on Reading Comprehension Models

EMNLP 2017 Outstanding Paper Award

Add an adversarial sentence to the paragraph to fool the model



Article: **Nikola Tesla**
Paragraph: "*In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.*"
Question: "*What city did Tesla move to in 1880?*"
Answer: *Prague*
Model Predicts: *Prague*



AddSent

What city did Tesla move to in 1880?

Prague

(Step 1) Mutate question

(Step 2) Generate fake answer

What city did Tadakatsu move to in 1881?

Chicago

(Step 3) Convert into statement

Tadakatsu moved the city of Chicago to in 1881.

(Step 4) Fix errors with crowdworkers, verify resulting sentences with other crowdworkers

Adversary Adds: ***Tadakatsu moved to the city of Chicago in 1881.***
Model Predicts: *Chicago*

Jia and Liang (2017): Adversarial Attacks on Reading Comprehension Models

EMNLP 2017 Outstanding Paper Award

- **Highly successful attacks**: over 16 models, F1 score drops from 75% to 36%
- **Their takeaway**: reading comprehension models are **overly stable**; unable to distinguish a sentence that answers the question from one that merely has words common with the question

Jia and Liang (2017): Adversarial Attacks on Reading Comprehension Models

EMNLP 2017 Outstanding Paper Award

- **Highly successful attacks**: over 16 models, F1 score drops from 75% to 36%
- **Their takeaway**: reading comprehension models are **overly stable**; unable to distinguish a sentence that answers the question from one that merely has words common with the question

Question for us: How does overstability manifest? Why do their attacks work?

Our contributions

- A workflow based on **attributions** (word-importances) to understand input-output behavior of networks
- Identify **weaknesses** in the networks as suggested by attributions
- Craft **adversarial examples** by **exploiting the weaknesses**
- **Explain** and **improve** Jia and Liang (2017)'s attacks

Attributions

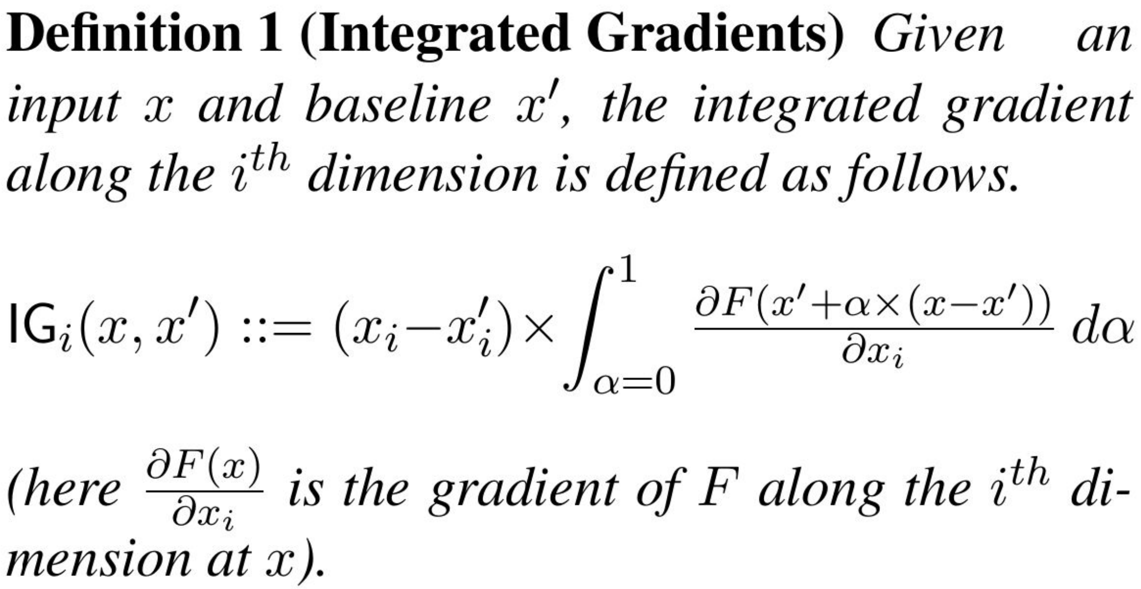
Problem statement: Attribute a complex deep network's prediction to input features, relative to a certain baseline (informationless) input

E.g. : attribute an object recognition network's prediction to its pixels, a text sentiment network's prediction to individual words

Explain `F(input) - F(baseline)` in terms of `input` features

Integrated Gradients

(Sundararajan et al (2017), ICML)



Definition 1 (Integrated Gradients) *Given an input x and baseline x', the integrated gradient along the i^{th} dimension is defined as follows.*

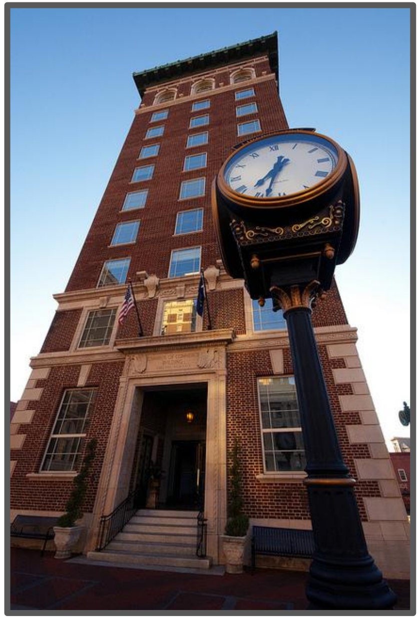
$$\mathsf{IG}_i(x, x') ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i}\, d\alpha$$

(here $\frac{\partial F(x)}{\partial x_i}$ is the gradient of F along the i^{th} dimension at x).

Why Integrated Gradients?

- Axiomatic justification (see Sundararajan et al (2017) for details)
- Ease of implementation; only gradient computations required
- running time < 0.5 seconds for a given input example

Visual QA attributions



Q: How symmetrical are the white bricks on either side of the building?
A: very

How symmetrical are the white bricks on either side of the building?

red: high attribution
blue: negative attribution
gray: near-zero attribution

Overstability

Drop all words from the dataset except ones which are frequently top attributions

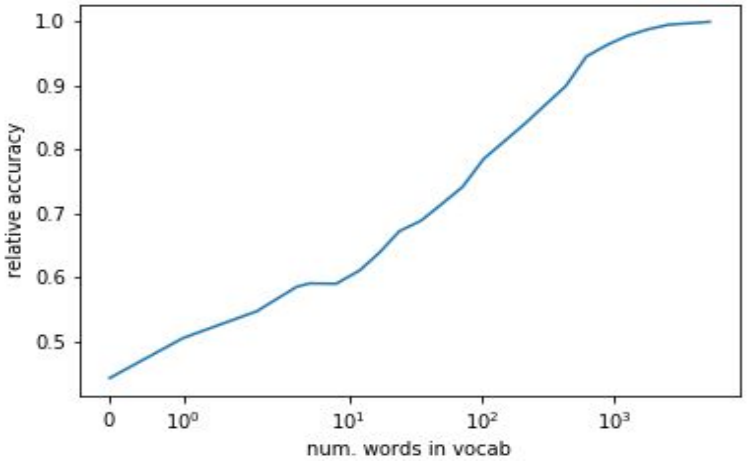
E.g. How many players scored more than 10 goals? → How many

Overstability

Drop all words from the dataset except ones which are frequently top attributions

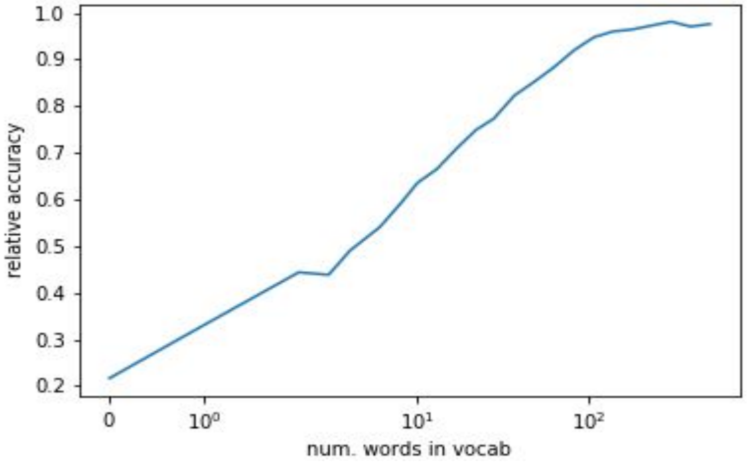
E.g. How many players scored more than 10 goals? → How many

Visual QA



color, many, what, how, doing, or, where, there, …

Neural Programmer



many, tm_token, how, number, total, after, …

Adversarial Examples

Stopword deletion attack

Delete contentless words from the question

show, tell, did, me, my, our, are, is, were, this, on, would, and, for, should, be, do, I, have, had, the, there, look, give, has, was, we, get, does, a, an, 's, that, by, based, in, of, bring, with, to, from, whole, being, been, want, wanted, as, can, see, doing, got, sorted, draw, listed, chart, only

Neural Programmer's accuracy falls from 33.5% to 28.5%

VQA model's accuracy falls from 61.1% to 52.0%

Subject ablation attack

Replace the subject of a question with a low-attribution noun from the vocabulary

Low-attribution nouns
'tweet',
'childhood',
'copyrights',
'mornings',
'disorder',
'importance',
'topless',
'critter',
'jumper',
'fits'

What is the **man** doing? ⟶ What is the **tweet** doing?
How many **children** are there? ⟶ How many **tweet** are there?

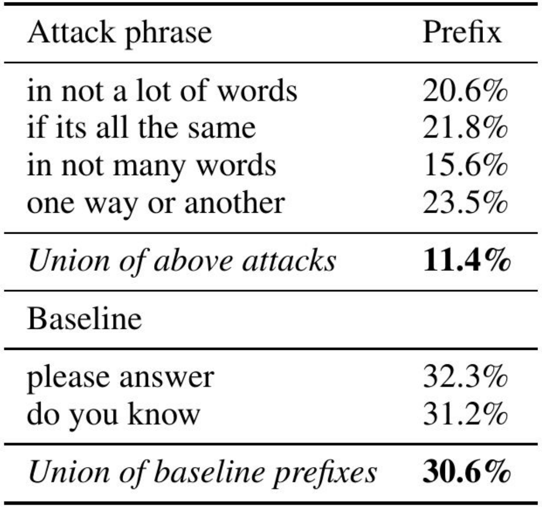
VQA model's response remains same 75.6% of the time on questions that it originally answered correctly

Question concatenation attacks

Prefix a content-free phrase to the question

Neural Programmer

Original accuracy: **33.5%**

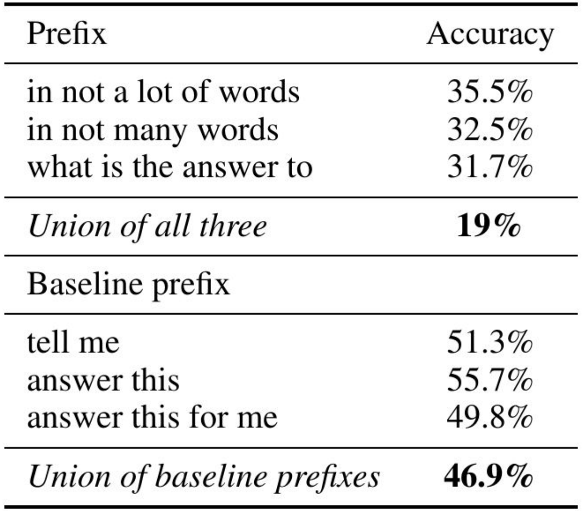


Attack phrase	Prefix
in not a lot of words	20.6%
if its all the same	21.8%
in not many words	15.6%
one way or another	23.5%
Union of above attacks	**11.4%**
Baseline	
please answer	32.3%
do you know	31.2%
Union of baseline prefixes	**30.6%**

Low attribution words

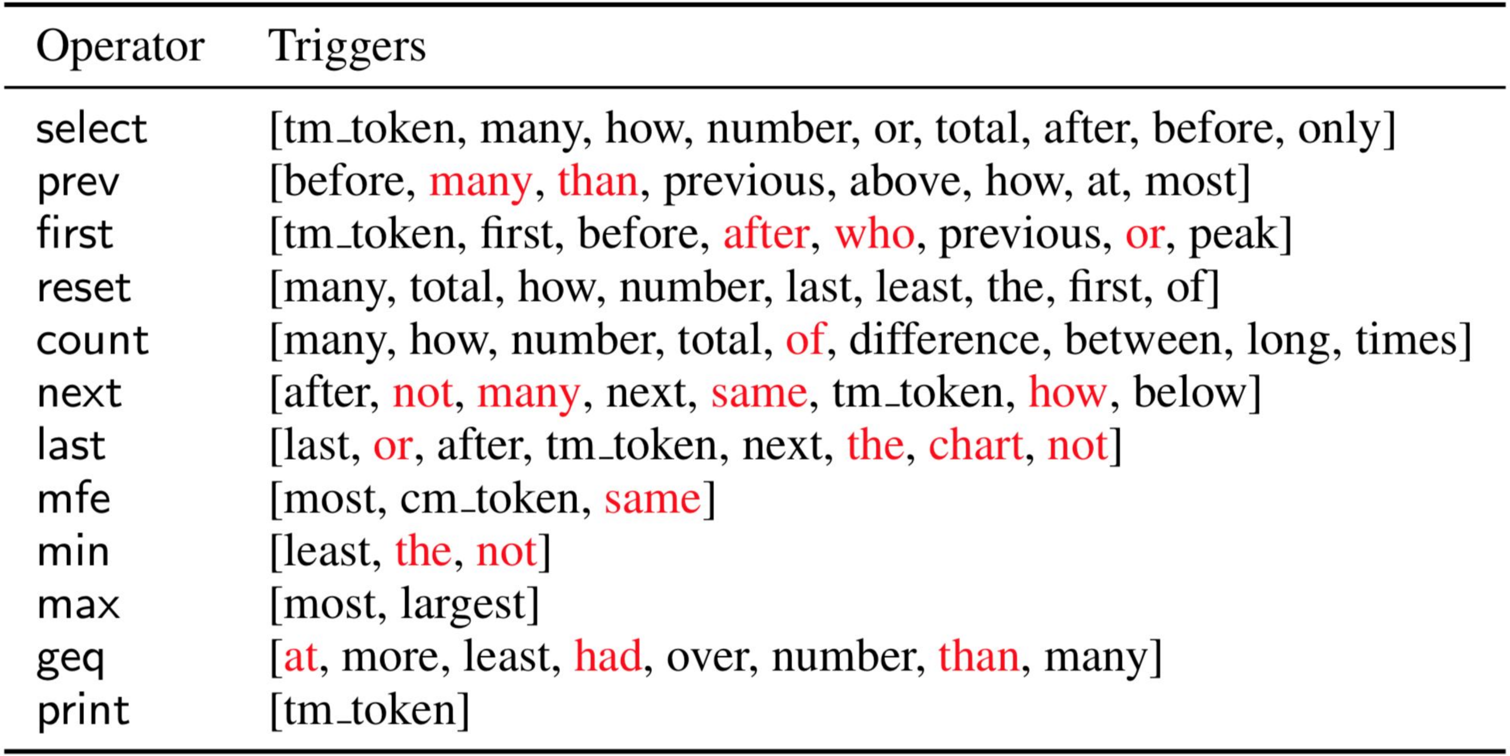
Visual QA

Original accuracy: **61.1%**



Prefix	Accuracy
in not a lot of words	35.5%
in not many words	32.5%
what is the answer to	31.7%
Union of all three	**19%**
Baseline prefix	
tell me	51.3%
answer this	55.7%
answer this for me	49.8%
Union of baseline prefixes	**46.9%**

Operator triggers in Neural Programmer



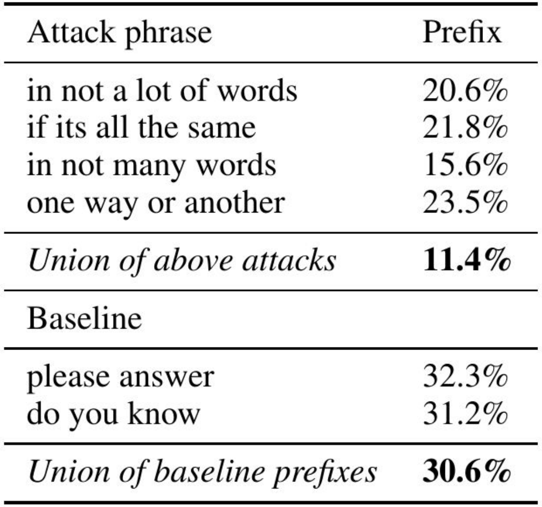
Operator	Triggers
select	[tm_token, many, how, number, or, total, after, before, only]
prev	[before, many, than, previous, above, how, at, most]
first	[tm_token, first, before, after, who, previous, or, peak]
reset	[many, total, how, number, last, least, the, first, of]
count	[many, how, number, total, of, difference, between, long, times]
next	[after, not, many, next, same, tm_token, how, below]
last	[last, or, after, tm_token, next, the, chart, not]
mfe	[most, cm_token, same]
min	[least, the, not]
max	[most, largest]
geq	[at, more, least, had, over, number, than, many]
print	[tm_token]

Question concatenation attacks

Prefix a content-free phrase to the question

Neural Programmer

Original accuracy: **33.5%**

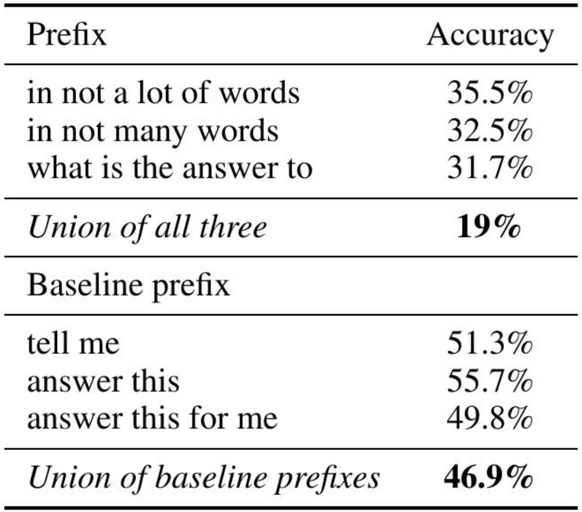


Attack phrase	Prefix
in not a lot of words	20.6%
if its all the same	21.8%
in not many words	15.6%
one way or another	23.5%
Union of above attacks	**11.4%**
Baseline	
please answer	32.3%
do you know	31.2%
Union of baseline prefixes	**30.6%**

Low-attribution words

Visual QA

Original accuracy: **61.1%**

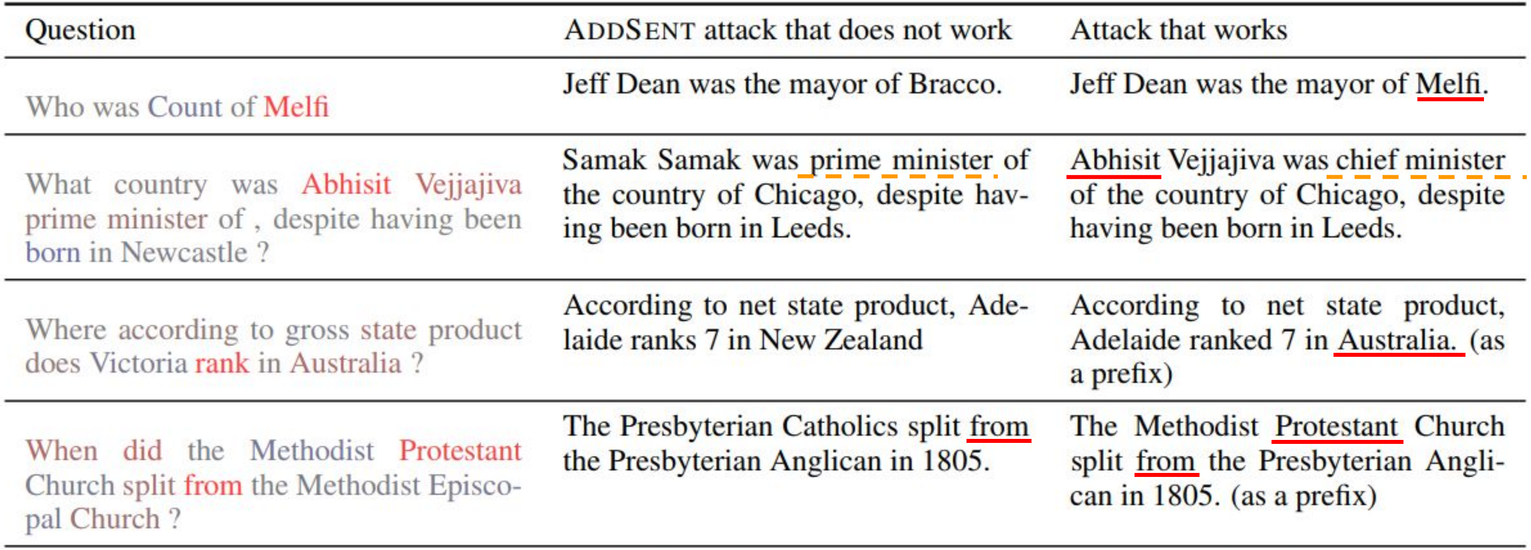


Prefix	Accuracy
in not a lot of words	35.5%
in not many words	32.5%
what is the answer to	31.7%
Union of all three	**19%**
Baseline prefix	
tell me	51.3%
answer this	55.7%
answer this for me	49.8%
Union of baseline prefixes	**46.9%**

Predicting the effectiveness of Jia and Liang (2017)'s adversarial attacks

Attacks are more likely to be effective when

- High-attribution words are **present** in the adversarial sentence
- Only low-attribution words are **mutated**



Question	ADDSENT attack that does not work	Attack that works
Who was Count of Melfi	Jeff Dean was the mayor of Bracco.	Jeff Dean was the mayor of Melfi.
What country was Abhisit Vejjajiva prime minister of , despite having been born in Newcastle ?	Samak Samak was prime minister of the country of Chicago, despite having been born in Leeds.	Abhisit Vejjajiva was chief minister of the country of Chicago, despite having been born in Leeds.
Where according to gross state product does Victoria rank in Australia ?	According to net state product, Adelaide ranks 7 in New Zealand	According to net state product, Adelaide ranked 7 in Australia. (as a prefix)
When did the Methodist Protestant Church split from the Methodist Episcopal Church ?	The Presbyterian Catholics split from the Presbyterian Anglican in 1805.	The Methodist Protestant Church split from the Presbyterian Anglican in 1805. (as a prefix)

red: high attribution, **blue**: negative attribution, **gray**: near-zero attribution

Summary

- An attribution-based workflow to look inside and understand weaknesses of a model
- Explained how overstability manifests - QA networks do not focus on the right words!
- Crafted adversarial examples and improved Jia and Liang (2017)'s attacks

Outlook

- Deep learning practitioners can **easily** use attributions to **look inside** models
- **Adding soft network constraints**
  - E.g. add bias to attention vector so as to limit the influence of “how”, “what”, etc.
- **Informed enrichment** of datasets
  - E.g. add more questions with word “symmetrical” such that answer is not “very”

If you would like to use our attribution-based workflow to understand your deep network/model

- https://**github**.com/**pramodkaushik**/acl18_results
- Contact me: pramodkm@uchicago.edu
- Ping me on Whova!

Thank you!