Interpretability & Security in NLP models

Dominika Basaj, 08/10/2018

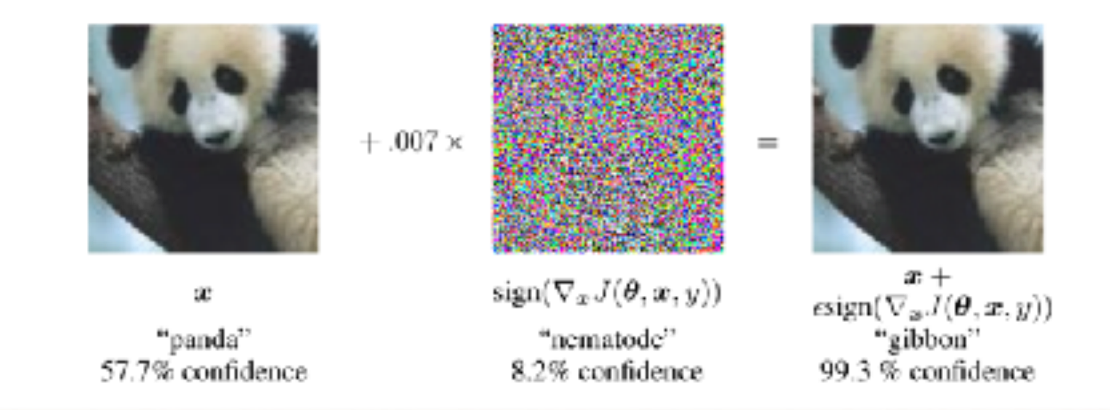
Agenda

- Introduction
- Related work
- Our research
- Future works

Adversarial examples

In general:

- Inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake
- First observed in 2014 by Szegedy et al. (2014)



Why does it matter?

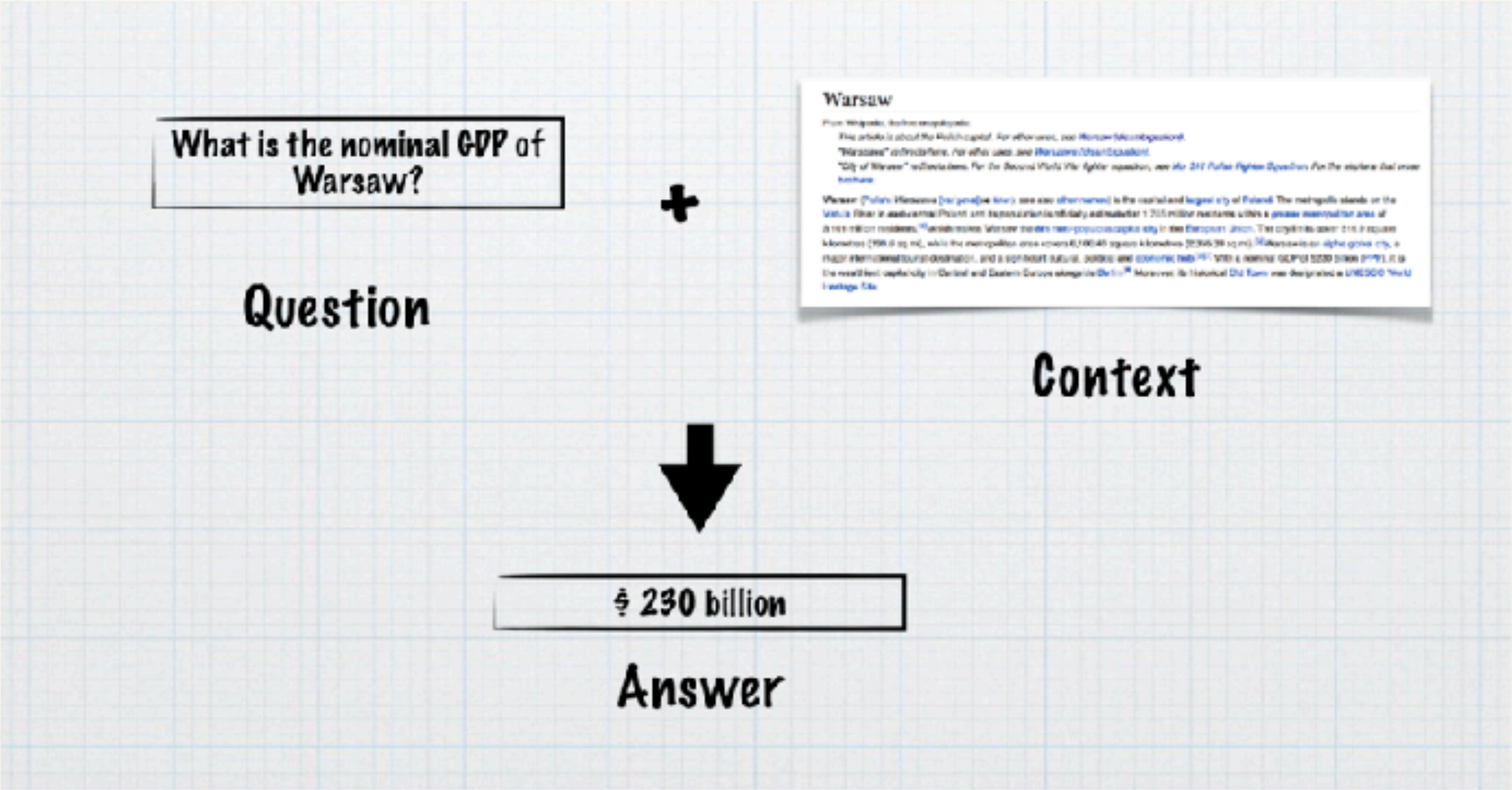
- There is growing recognition that ML exposes new vulnerabilities in software systems
- A good aspect of security to work on
- They represent a concrete problem in AI safety
- Fixing them is difficult enough that it requires a serious research effort

Source: https://blog.openai.com/adversarial-example-research/, NIPS 2018 Workshop on Security in Machine Learning

Szegedy, Christian, Zaremba, Wojciech, Sutskever, Ilya, Bruna, Joan, Erhan, Dumitru, Goodfellow, Ian J., and Fergus, Rob. Intriguing properties of neural networks. ICLR

Goodfellow, Ian , Schlens, Johnatan, Szegedy, Christian, . Explaining and harnessing adversarial examples. 2015

Recap on machine comprehension systems



Popular SQuAD dataset

Adversarial examples in NLP (1)

Adversarial Examples for Evaluating Reading Comprehension Systems, Jia Robert, Liang Percy, EMNLP 2017

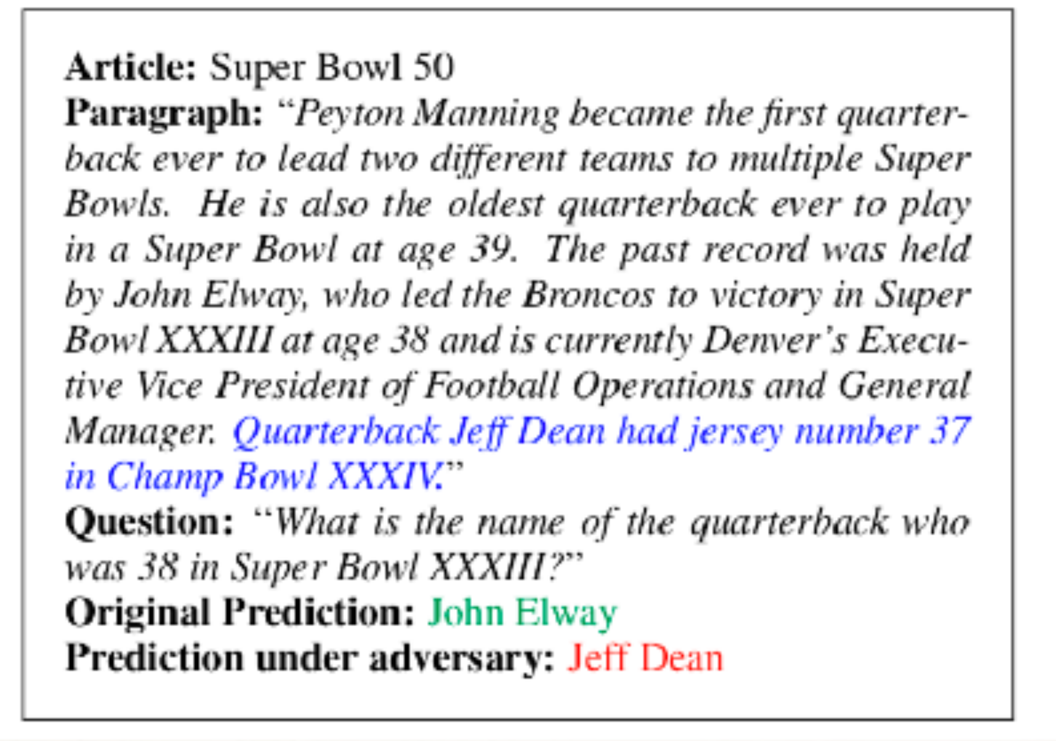
- Do systems understand language?
- Insert adversarial sentences to distract computer systems without changing the correct answer

Adversarial techniques used in paper:

1. AddSent - generate sentences that look similar to the question, but do not actually contradict the correct answer
2. AddAny - any sequence of d words, regardless of grammaticality
3. AddCommon - like ADDANY except it only adds common words

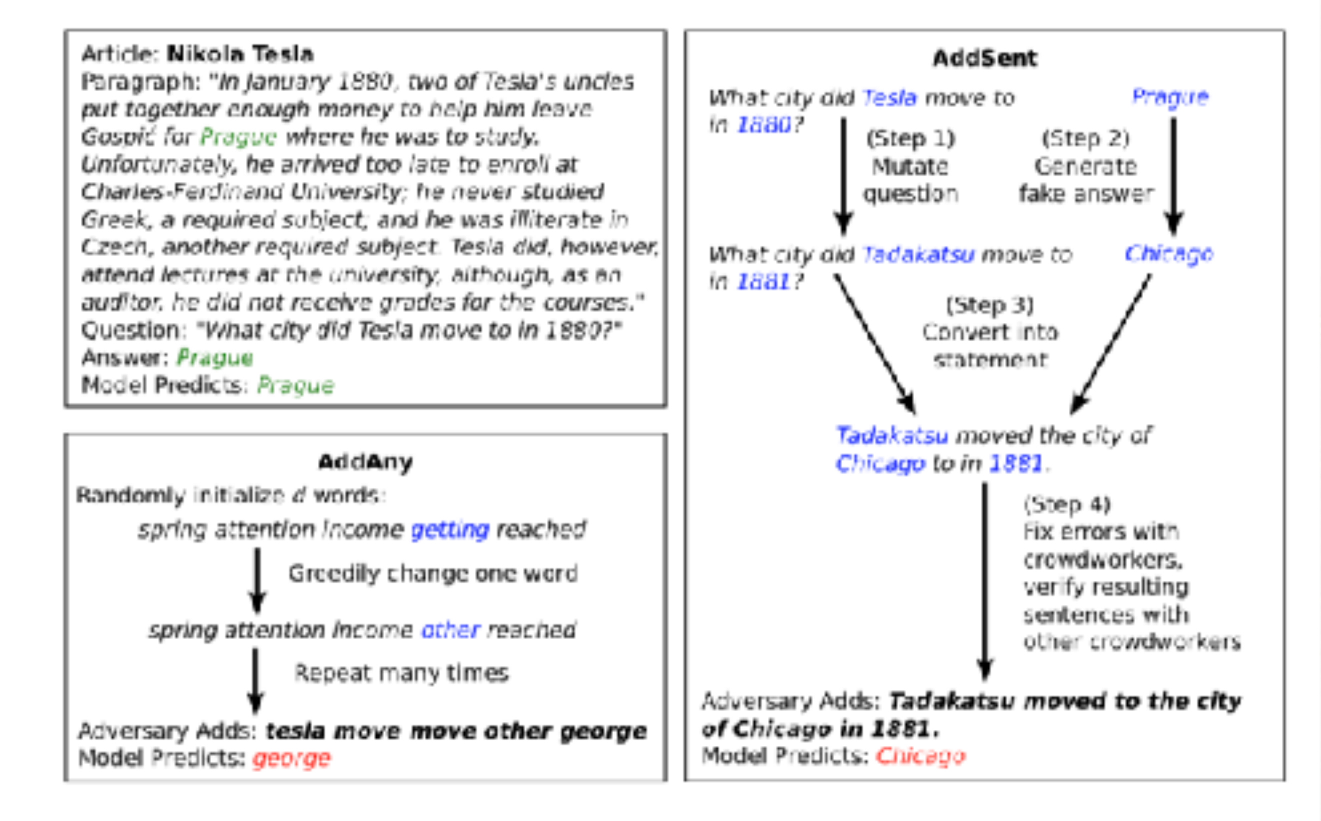
Overstability:
the inability of a model to distinguish a sentence that actually answers the question from one that merely has words in common with it

Adversarial examples in NLP (1.1)



Article: Super Bowl 50
Paragraph: "*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. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Example of BiDAF model being fooled.



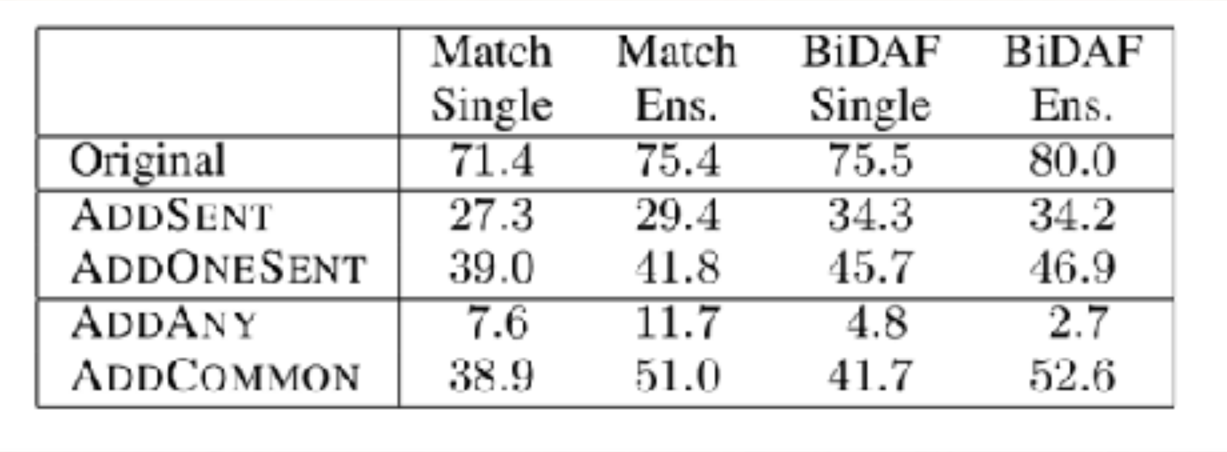
Article: **Nikola Tesla**
Paragraph: "*In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate 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*

AddAny
Randomly initialize *d* words:
spring attention income getting reached
Greedily change one word
spring attention income other reached
Repeat many times
Adversary Adds: ***tesla move move other george***
Model Predicts: *george*

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*

Automatic creation of adversarial examples.

Adversarial examples in NLP (1.2)



	Match Single	Match Ens.	BiDAF Single	BiDAF Ens.
Original	71.4	75.4	75.5	80.0
AddSent	27.3	29.4	34.3	34.2
AddOneSent	39.0	41.8	45.7	46.9
AddAny	7.6	11.7	4.8	2.7
AddCommon	38.9	51.0	41.7	52.6

We can observe that adversarial techniques reduce accuracy of each state-of-the-art model.

„accuracy of sixteen published models drops from an average of 75% F1 score to 36%; when the adversary is allowed to add ungrammatical sequences of words, average accuracy on four models decreases further to 7%"

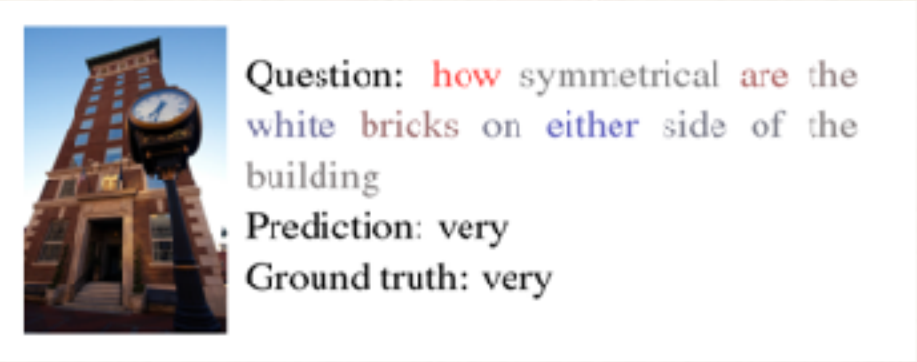
Adversarial examples in NLP (2)

Did the model understand the question?, Mudrakarta Pramod, Taly Ankur, Sundararajan Mukund, Dhamdhere Kedar, ACL 2018

Model: 3 Q&A tasks-tabular, text and image data

- Uses integrated gradients (Sundararajan et al, 2017) technique to understand which words in questions have high attributions to model outcome
- Overstability for questions in Q&A systems
- Overall, informative words in the question (e.g., nouns) often receive very low attribution

"how spherical are the white bricks on either side of the building",
"how soon are the bricks fading on either side of the building",
"how fast are the bricks speaking on either side of the building"

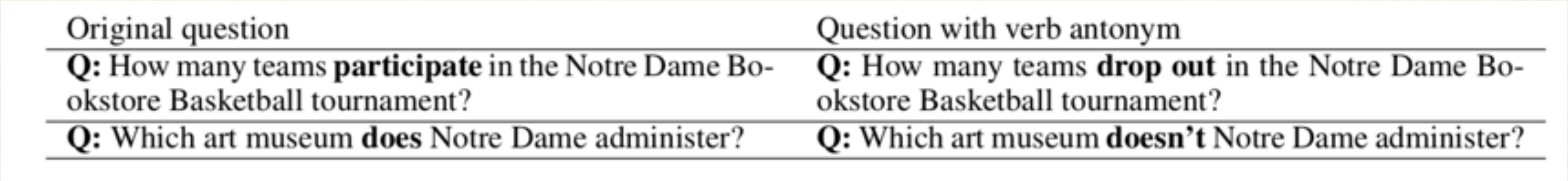


Question: how symmetrical are the white bricks on either side of the building
Prediction: very
Ground truth: very

Our work (1)

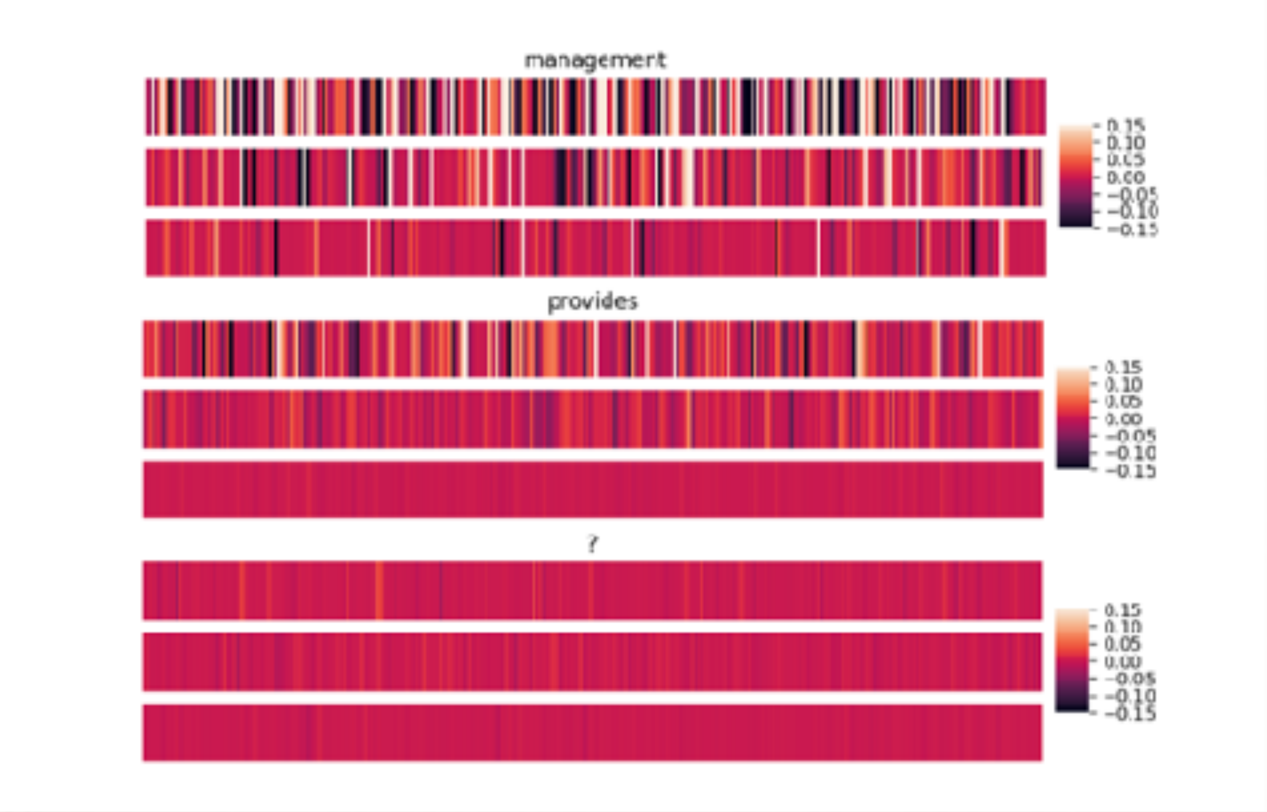
Does it care what you asked? Understanding Importance of Verbs in Deep Learning QA System

- We inspect importance of verbs in a deep learning QA system trained on SQuAD dataset
- Applied adversaries:
  - swap verbs to their negations (based on WordNet)
  - analyze how many % changed prediction
- No change of answer - 90.5% of questions
- We analyze question attention scores and entropy/variance of values of neurons in hidden layers

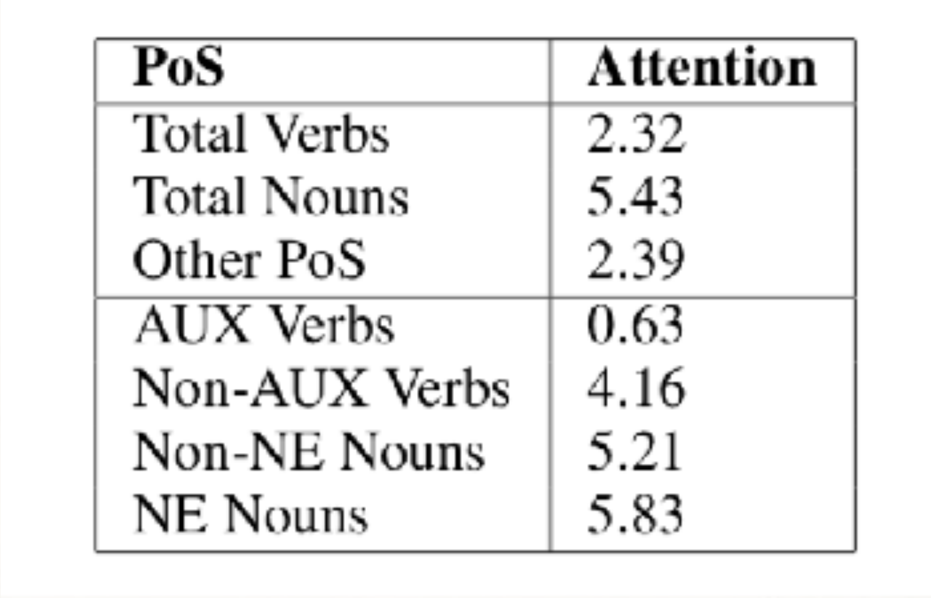


Original question	Question with verb antonym
Q: How many teams **participate** in the Notre Dame Bookstore Basketball tournament?	**Q:** How many teams **drop out** in the Notre Dame Bookstore Basketball tournament?
Q: Which art museum **does** Notre Dame administer?	**Q:** Which art museum **doesn't** Notre Dame administer?

Our work (1.1)



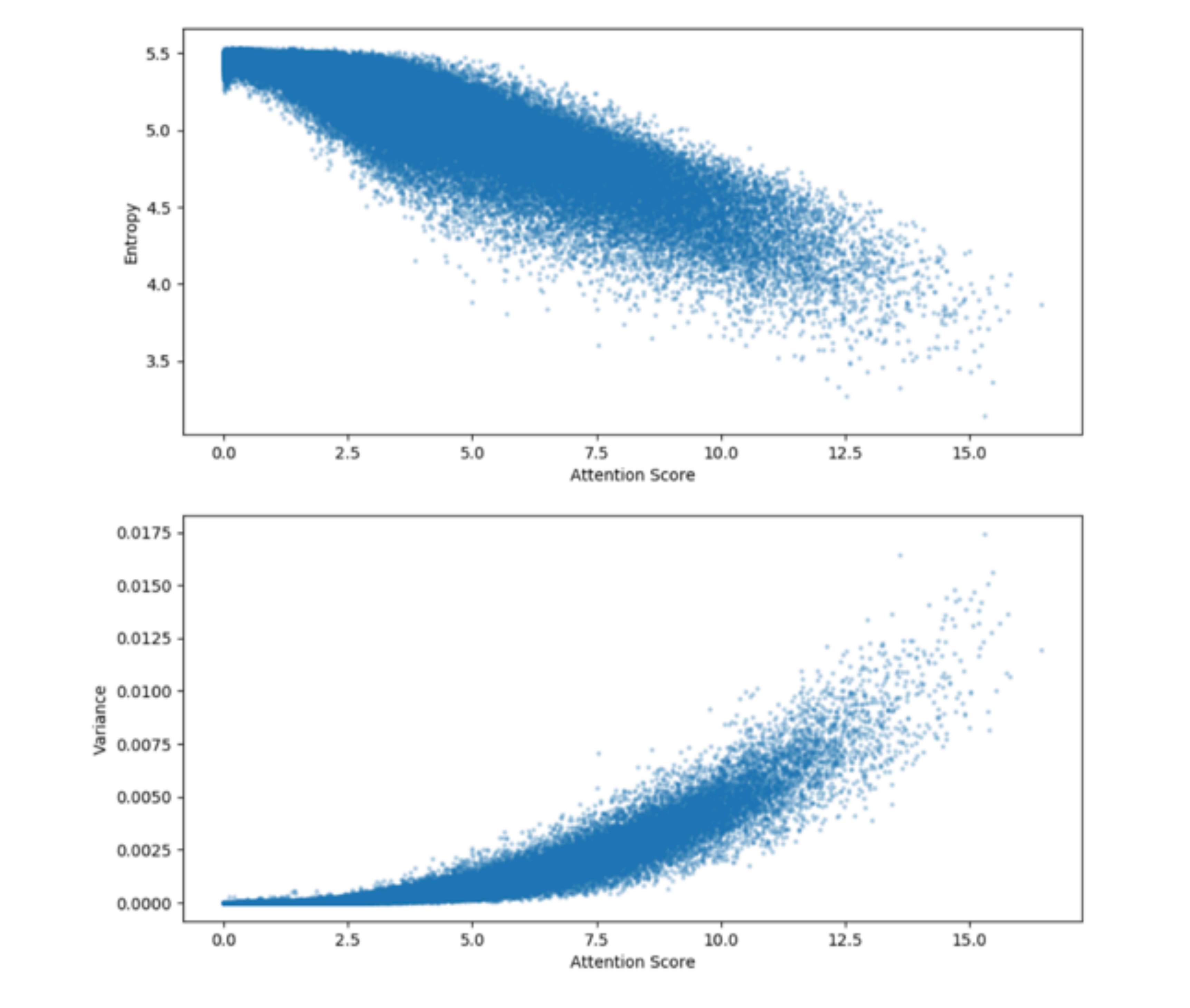
Visualization of values in LSTM hidden layers



PoS	Attention
Total Verbs	2.32
Total Nouns	5.43
Other PoS	2.39
AUX Verbs	0.63
Non-AUX Verbs	4.16
Non-NE Nouns	5.21
NE Nouns	5.83

Attention scores per part-of-speech

Our work (1.2)

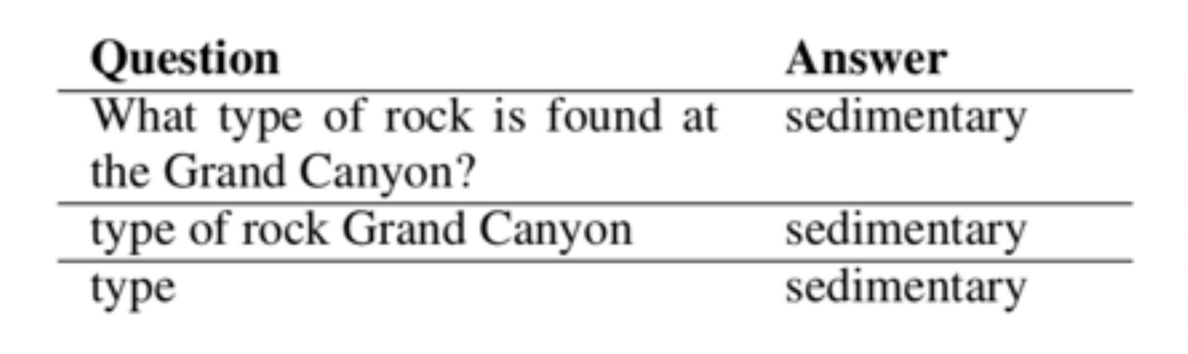


Scatter plot of entropy / variance (y-axis) and attention (x-axis)

Our work (2)

How much should you ask? On the question structure in QA systems

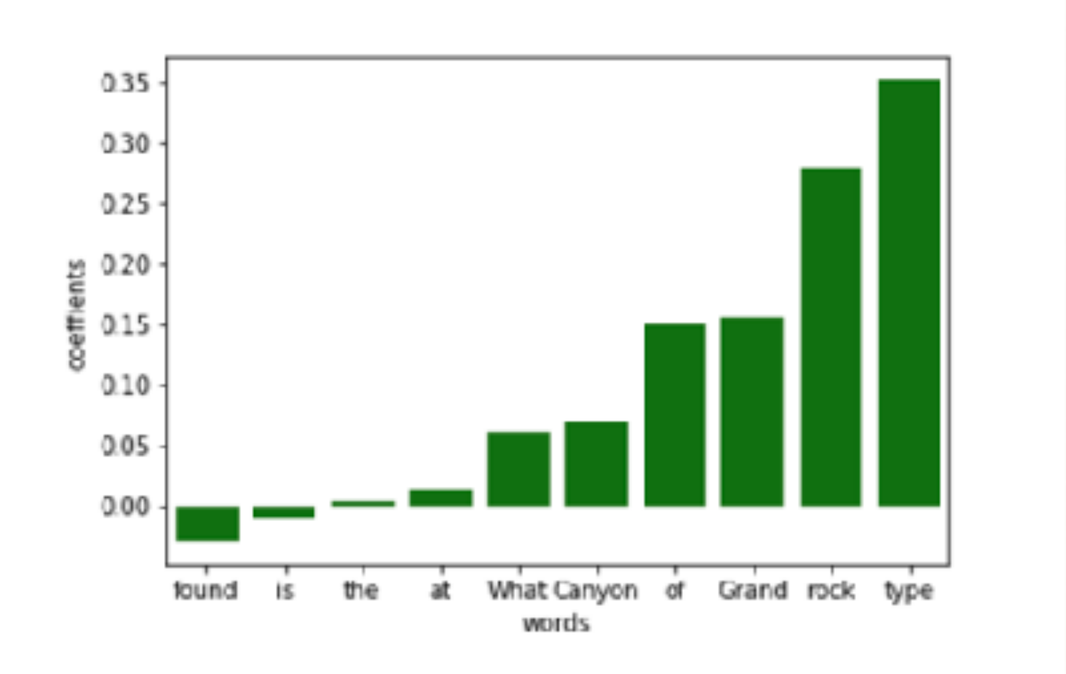
- Used LIME for detection of important words in questions
- Conclusion: grammar and natural language is disregarded by the model
- QA system returns true answer once we type in just selected keywords without keeping the sentence structure



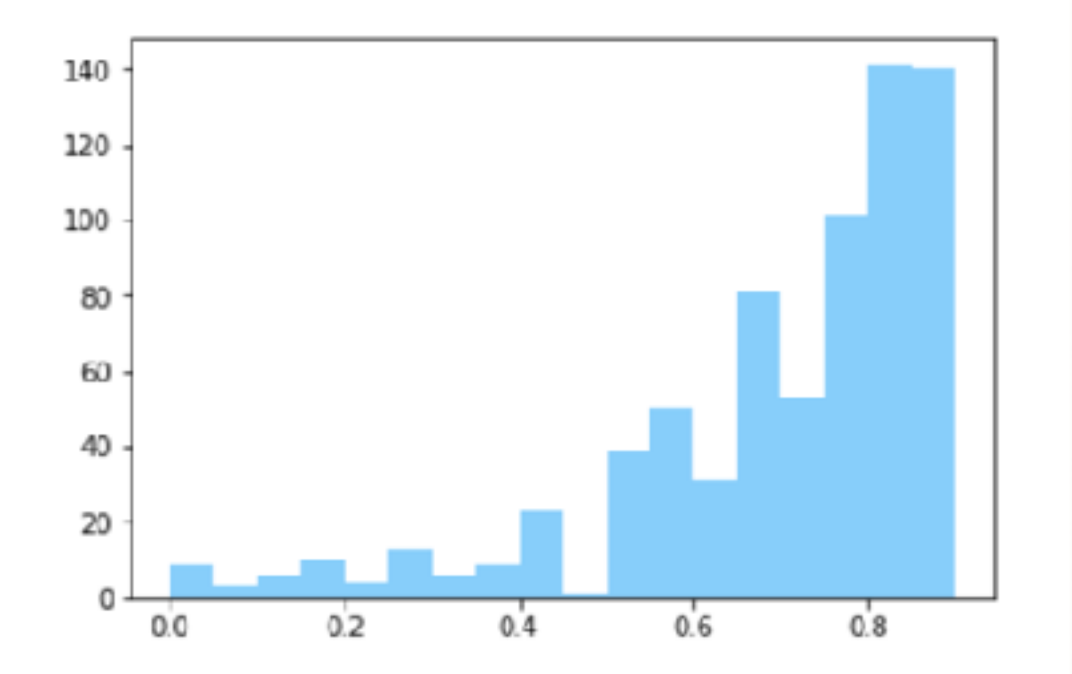
Question	Answer
What type of rock is found at the Grand Canyon?	sedimentary
type of rock Grand Canyon	sedimentary
type	sedimentary

Questions and answers after removing subsequent words

Our work (2.1)



Lime coefficients per word



Percentage of words that must be removed in order to change answer - distribution