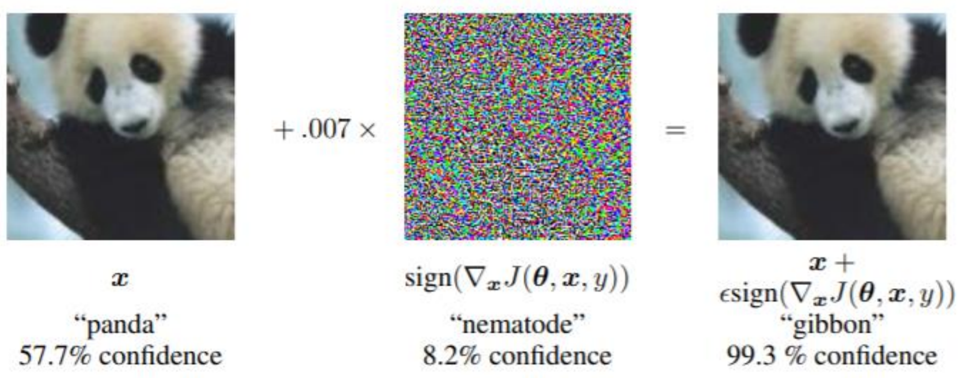
Adversarial Attack to Semantic Parsers

Weiliang Tang, Shilin He (T.A.), Michael Lyu (Prof.)

Introduction of New Adversarial Task for Semantic Parser

Previous Work

- **Adversarial attack to image classification models** [Goodfellow et al., 2014]



- **Adversarial attack to text classification models** [Ebrahimi et al., 2017]

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mood of optimism. 57% **World**

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mooP of optimism. 95% **Sci/Tech**

The Semantic Parser

Natural Language ⇒ Logical Form

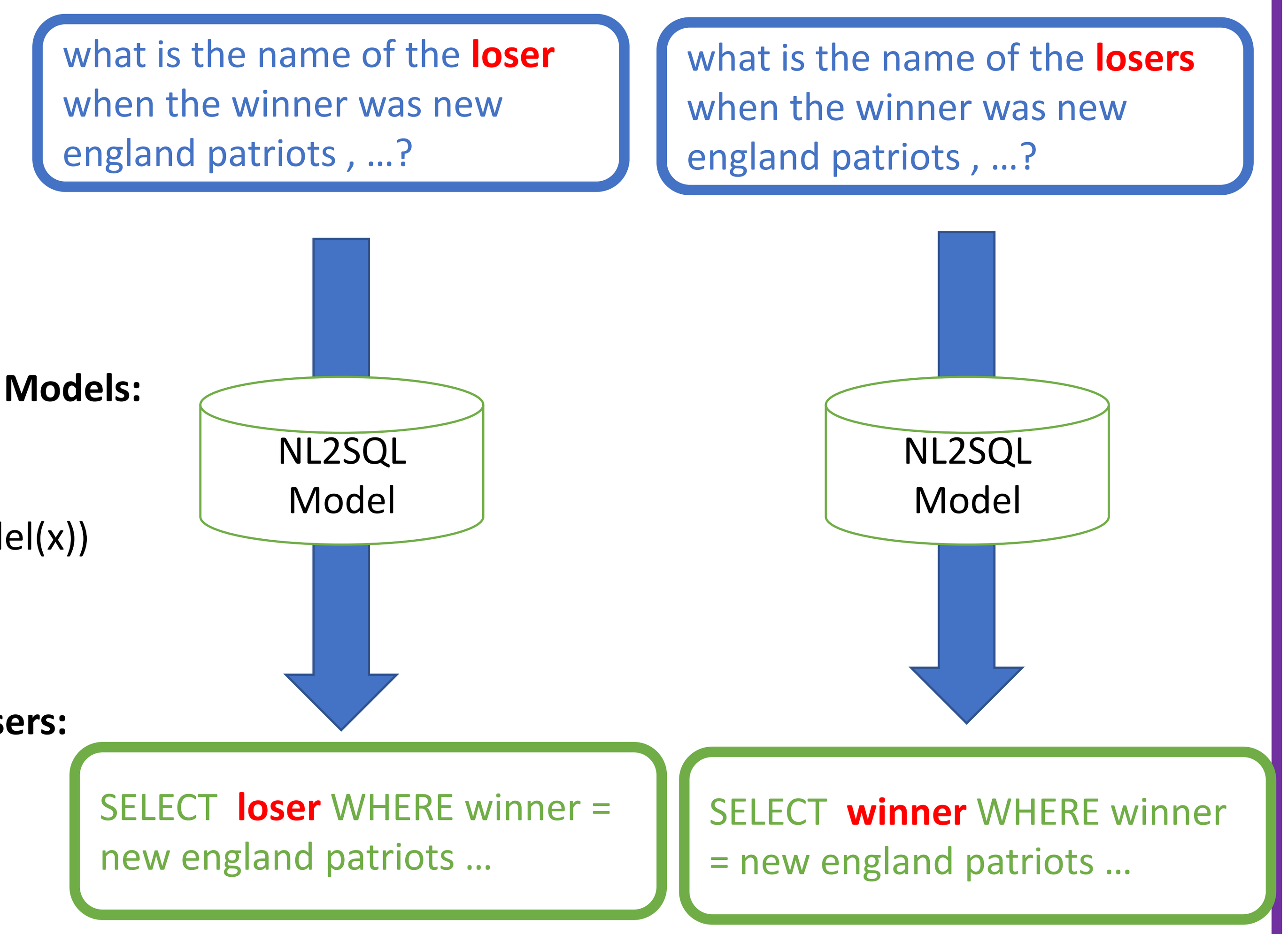
Adversarial Attack to Text Classification Models:

Generate x* ,where

- Semantic(x*) = Semantic(x)
- Semantic(Model(x*)) ≠Semantic(Model(x))

New Challenges to Attack Semantic Parsers:

- The input is short, change of input is visually distinguishable.
- The input space is discrete.



Generating Adversarial Examples

Evaluation Metric:

- **correct ratio**: correct predictions/ input data
- **diff ratio**: diff. predictions/ perturbed input data
- **valid ratio**: predictions which keep the semantic meaning unchanged / diff. outputs

Basic Method: **Fast Gradient Method (FGM)**

Algorithm:

FGM

$$
\begin{array}{l}
\text{// } grad\_data = (\text{input\_len} \times \text{embedding\_size}) \\
\textbf{for } i = 0 \textbf{ to } length[grad\_data] - 1 \\
\qquad word\_grad[i] = \|grad\_data[i]\| \\
target\_word = \arg\max(word\_grad) \\
perturbed\_word = \underset{w \in embed\_space}{\arg\min} \|word[idx] + \epsilon \cdot grad\_data[idx] - w\|
\end{array}
$$

Experiment Settings:

- **Model**: Coarse2Fine (Accuracy: 71.7%) [Dong and Lapata, 2018]
- **Dataset**: WikiSQL [Zhong et al., 2017]
- **Reimplement accuracy**: 67.46%

Experiment Result:

- **The method is able to generate adversarial examples**

What is the air **force** cross when … => SELECT **airforcecross** WHERE…
What is the air **forces** cross when … => SELECT **navyforcecross** WHERE…

it has a fa cup goals smaller than 4, what is the total number of total **apps** ? => SELECT MAX **totalapps** facupgoals = 4
it has a fa cup goals smaller than 4, what is the total number of total **app** ? => SELECT MAX **facupapps** facupgoals = 4

Fast Gradient Method

what **gender** is quentin ? => SELECT **gender** WHERE name = quentin
what **genders** is quentin ? => SELECT **status** WHERE name = quentin

How many **types** of organization … => SELECT MAX **types** WHERE…
How many **kinds** of organization … => SELECT MAX **organization** WHERE…

- **The larger the ε is, the higher diff ratio and lower valid ratio it will be when ε is relatively small**

- **Cause: Under fitting problem:** Some words are crowded in small area in embedding space, the word untrained is easily been misguided by the trained words next to it .
- **Drawback**: The choice of word neglects the semantic environment, one word can be perturbed only into another fixed word under no circumstances.

Improvement Using Bert

Algorithm:

Bert-FGM

$$
\begin{array}{l}
\textbf{for } i = 0 \textbf{ to } 3 \\
\qquad \text{// } grad\_data = (\text{input\_len} \times \text{embedding\_size}) \\
\qquad \textbf{for } i = 0 \textbf{ to } length[grad\_data] - 1 \\
\qquad\qquad word\_grad[i] = \|grad\_data[i]\| \\
\qquad target\_word\_list = n\_\arg\max(word\_grad) \\
\qquad \textbf{for } i = 0 \textbf{ to } length[idx\_list] - 1 \\
\qquad\qquad target\_word = target\_word\_list[i] \\
\qquad\qquad bert\_list = Bert(sen, target\_word, 10) \\
\qquad\qquad word\_list = \underset{w \in bert\_list}{\arg\max}\ c \cdot bert\_prob[w] + cos\_simi(\epsilon \cdot grad\_data[idx], w - target\_word) \\
\qquad perturbed\_word = \underset{w \in word\_list}{\arg\max}\ c \cdot bert\_prob[w] + cos\_simi(\epsilon \cdot grad\_data[idx], w - target\_word) \\
\qquad word \Rightarrow perturbed\_word
\end{array}
$$

Experiment Result:

- **Successful examples are in more various forms**
- **Examples are of more semantic consistency**

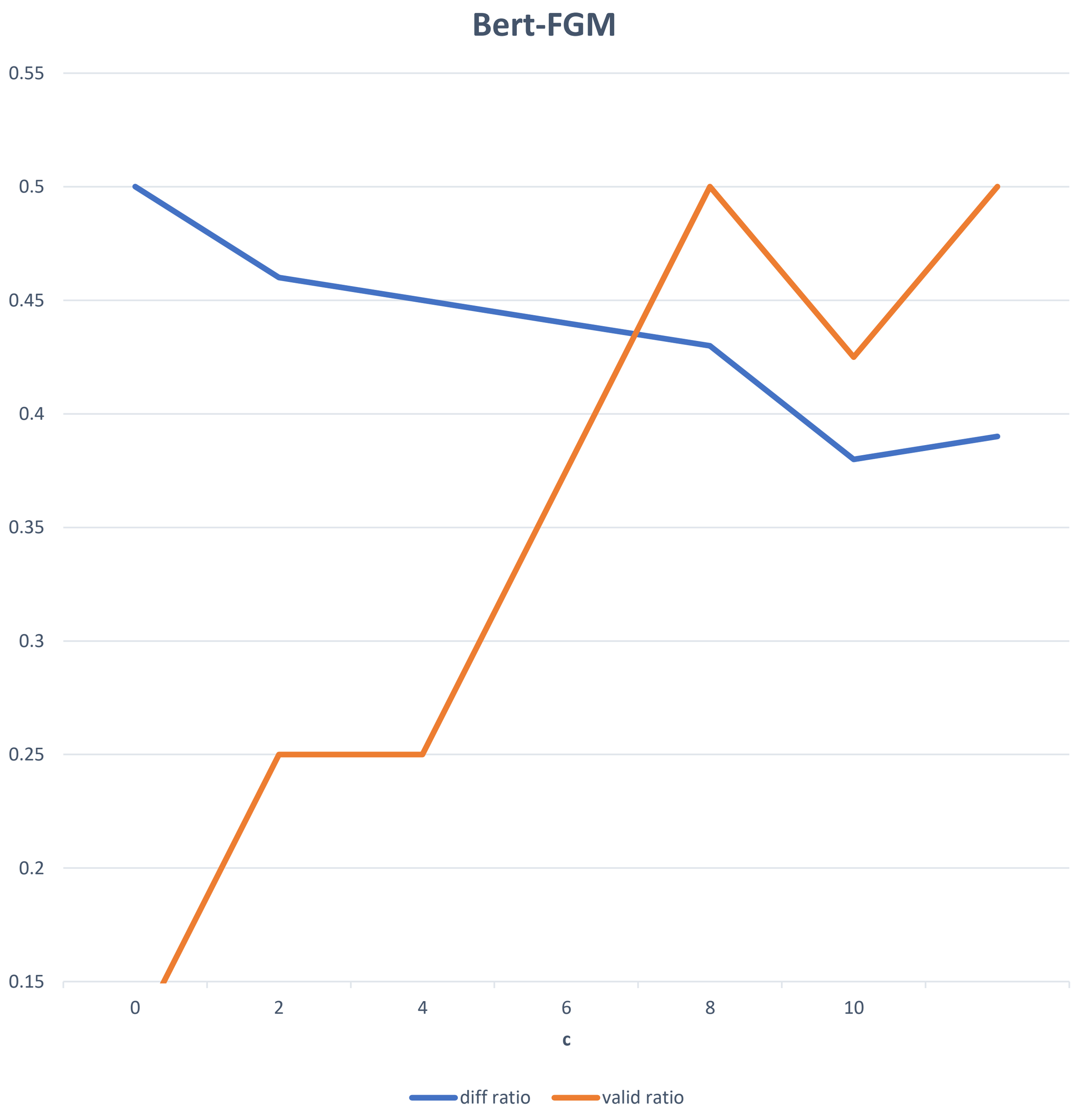
what is height , **when** rank is less than 20… => SELECT height WHERE built = 2005 AND **name** = the edge -lrb- c -rrb-
What is height, **where** rank is less than 20… => SELECT height WHERE built = 2005 AND **rank** = the edge -lrb- c -rrb-

which athlete 's rank is more than 15 when the result is less than 7.68 **,** the group is b , and the nationality listed **is** great britain ? => SELECT athlete WHERE group = b AND **nationality** = great britain AND rank = 15 AND result = 7.68
which athlete 's rank is more than 15 when the result is less than 7.68 **and** the group is b , and the nationality listed **in** great britain ? => SELECT athlete WHERE group = b AND **group** = great britain AND rank = 15 AND result = 7.68

what is the smallest period -lrb- days -rrb- to have a planetary mass **of** 1, and … => SELECT MIN period-lrb-days-rrb- WHERE **planetarymass**-lrb-m = 1 …
what is the smallest period -lrb- days -rrb- to have a planetary mass **at** 1, and … => SELECT MIN period-lrb-days-rrb- WHERE **stellarmass**-lrb-m = 1…

what is type , … and **when** etymology **is** son of jens ? => SELECT type WHERE **etymology** = son of jens
what is type , … and **whose** etymology **are** son of jens ? => SELECT type WHERE **surname** = son of jens

- **A trade off between diff ratio and valid ratio**
  - The smaller the c is, the word is more likely to follow the gradient straightly, the higher diff ratio is.
  - The bigger the c is , the word is more likely to make sense, the higher valid ratio is.



- **Our method successfully elaborates the valid ratio compared to previous simple FGM method**

- [Goodfellow et al., 2014] Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572.*
- [Moosavi-Dezfooli et al., 2016] Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582.
- [Zhong et al., 2017] Zhong, V., Xiong, C., and Socher, R. (2017). Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.
- [Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Reference]