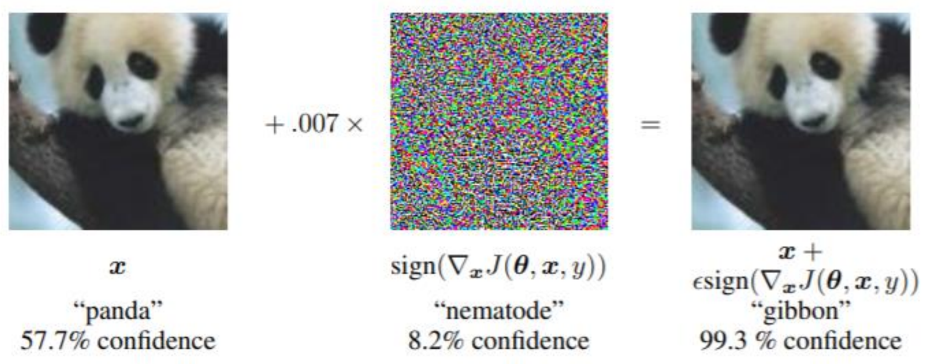
Adversarial attack to Semantic Parser

Weiliang Tang, Shilin He (TA) , Michael Lyu (Prof.)

Introduction of New Adversarial Task for Semantic Parser

Adversarial attack to image classification model



Adversarial attack to text classification model

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

New Challenge to Attack Semantic Parser:

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

Semantic parser:

Natural Language ⇒ Logical Form

A new definition of adversarial example: x* where

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

Visual adversarial ⇒ Semantic adversarial

Long input
Continuous input

Short discrete input
Semantic task

what is the name of the **loser** when the winner was new england patriots , …?

NL2SQL Model

SELECT **loser** WHERE winner cond_op new england patriots …

what is the name of the **losers** when the winner was new england patriots , …?

Same NL2SQL Model

SELECT **winner** WHERE winner cond_op new england patriots …

Generating Adversarial Examples

Measurement:
Correct ratio: correct predictions/ input data
Differ ratio: diff predictions/ perturbed input data.
Valid ratio: predictions which keep the semantic meaning unchanged / different outputs.

Basic Method: **Fast Gradient Method**
Algorithm:

FGM

```
// grad_data = (input_len × embedding_size)
for i = 0 to length[grad_data] − 1
        word_grad[i] = ||grad_data[i]||
target_word = arg max(word_grad)
perturbed_word = arg min_{w∈embed_space} ||word[idx] + ε · grad_data[idx] − w||
```

Experiment Result:

- The larger the ε is, the higher diff ratio and lower valid ratio it will be when ε is relatively small.
- Some pattern is shown in the successful perturbed examples:

1. Among all the successful example, 36% is done by changing a word in single form to plural form.

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

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

2. Substitute the word with its synonym

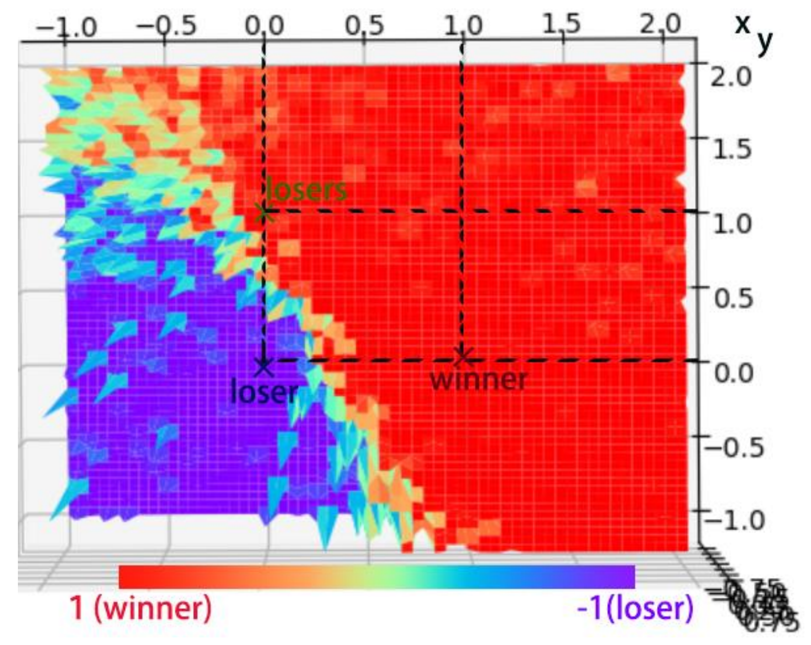
How many types of organization … => SELECT MAX **types** WHERE…
How many kinds of organization … => SELECT MAX **organization** WHERE…

- **Drawback**:
  The choice of word neglects the semantic environment around it, one word can be perturbed only into another fixed word under on circumstances

Fast Gradient Method

Reason: Under-fitting problem in NL2SQL task

The distribution of z (see below) on a plane in the high dimensional space



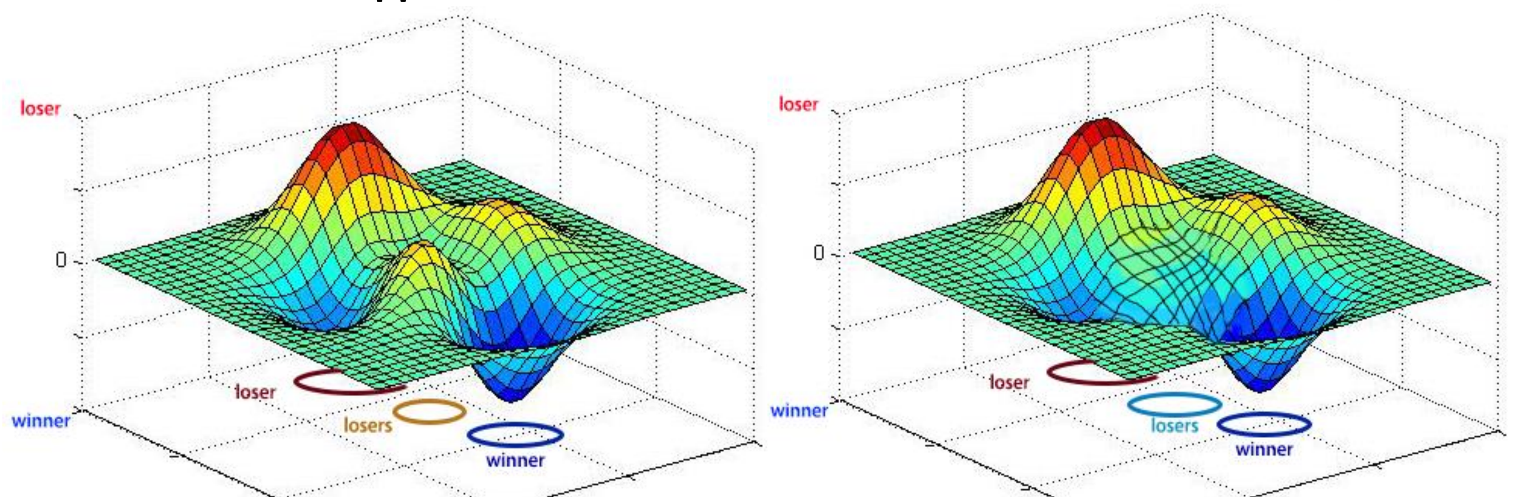
$$z = \frac{(p(sel\_op='loser') - p(sel\_op='winner'))}{(p(sel\_op='loser') + p(sel\_op='winner'))}$$

$$p(sel\_op) = Model(\vec{v}('losers') + x \cdot (\vec{v}('losers') - \vec{v}('loser')) + y \cdot (\vec{v}('winner') - \vec{v}('loser')); \theta)$$

The plane where "loser", "losers" and "winner" lie in

$$\text{when } x = 1, y = 0,\ p(sel\_op) = Model(\vec{v}('losers'))$$
$$\text{when } x = 0, y = 1,\ p(sel\_op) = Model(\vec{v}('winner'))$$
$$\text{when } x = y = 0,\ p(sel\_op) = Model(\vec{v}('loser'))$$

What's supposed to be | **What it looks like actually**



Under fitting problem: some words are crowded in a small area, the word untrained is easily been misguided by the trained words around it

- **New adversarial feature for NL2SQL model:** The header of SQL usually are of the same type and sometimes very close to each other, the header can be vulnerable under adversarial attack

Improvement Using Bert

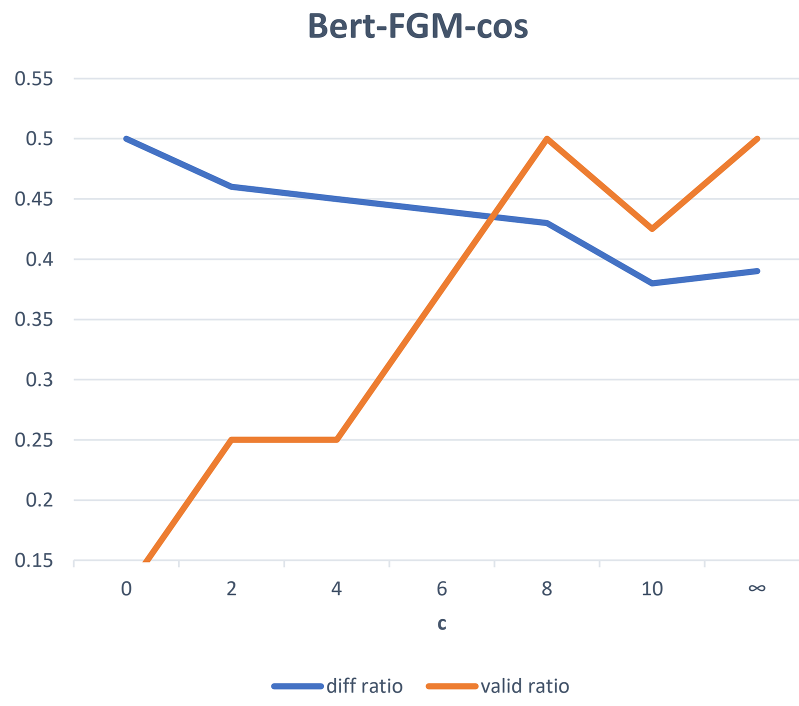
Algorithm:

Bert-FGM

```
for i = 0 to 3
        // grad_data = (input_len × embedding_size)
        for i = 0 to length[grad_data] − 1
                word_grad[i] = ||grad_data[i]||
        target_word_list = n_arg max(word_grad)
        for i = 0 to length[idx_list] − 1
                target_word = target_word_list[i]
                bert_list = Bert(sen, target_word, 10)
                word_list = arg max_{w∈bert_list} c · bert_prob[w] + cos_simi(ε · grad_data[idx], w − target_word)
        perturbed_word = arg max_{w∈word_list} c · bert_prob[w] + cos_simi(ε · grad_data[idx], w − target_word)
        word ⇒ perturbed_word
```

Experiment result:

- A trade off between diff ratio and valid ratio
  - The smaller the c is, the more dominant the cosine_similarity will be, the word is more likely to follow the gradient straightly, the higher diff ratio is.
  - The bigger the c is , the more dominant the bert_probability will be, the word is more likely to make sense, the higher valid ratio is, but it may not follow the gradient too much.
- This method successfully elaborate the valid ratio compared to previous simple FGM method

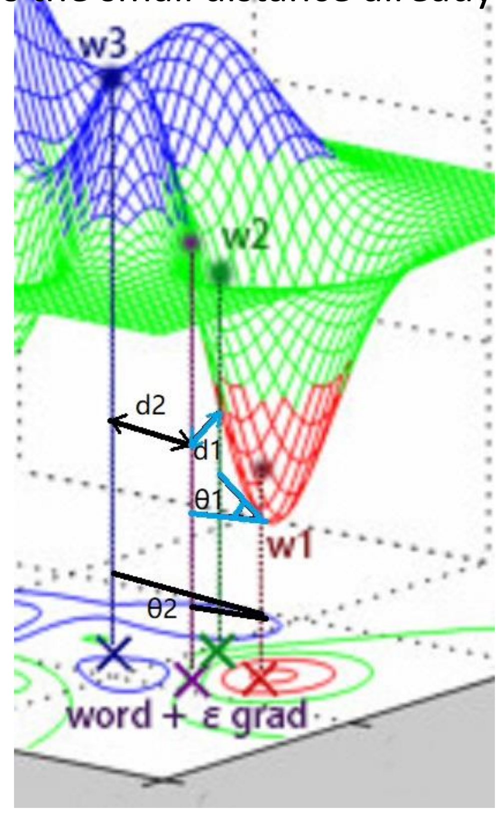


- A more variety of forms of successful example is shown
- A more semantic consistency is shown after substitution

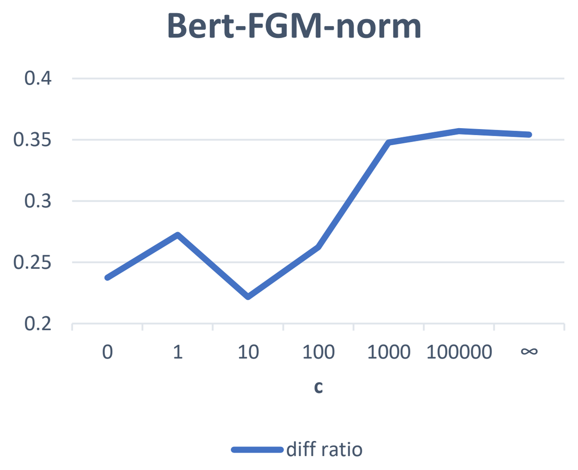
Original sentence	Perturbed sentence
what is height , **when** rank is less than 20…	What is height, **where** rank is less than 20…
When total goals have a fa cup apps larger than 1 ,…,what is the total number?	**if** total goals have a fa cup apps larger than 1 ,…,what is the total number?
what is the smallest period -lrb- days -rrb- to have a planetary mass **of** 1, and …	what is the smallest period -lrb- days -rrb- to have a planetary mass **at** 1, and …

Cosine similarity is a more reasonable choice：

- d1 < d2
  loss(d2) > loss(d1)
- cos_similarity describe the degree of following the gradient better since Bert ensures the small distance already



- Unreasonable result occurs if using norm distance



Weiliang Tang
Email:1155107670@link.cuhk.edu.hk