# Cross-Domain Transferability of Adversarial Examples in NLP

David Guzmán  Jean Charle Yaacoub

## Introduction

- Existing attacks rely directly or indirectly on the source training data, which hampers their transferability to other domains.
- Previous work in CV [1] has explored the idea of the transferability of adversarial examples across domains with compelling results that demonstrated such domain invariant adversaries.
- A true **black-box** attack must be able to fool models across different target domains/tasks without ever being explicitly trained on those target domains/tasks.
- In this work we explore this idea in NLP by exploring different domain and task settings and looking at how adversarial examples transfer across them.
- To the best of our knowledge, the work closest to ours is [2], which only explores the **similar domain, same task** setting.

## Methodology

- Pre-trained models were chosen from HuggingFace [3, 4].
- Models were selected based on architecture, domain and task, as showed in table 1.
- Domain is defined as the data distribution the model is trained on.
- Settings start from the **similar domain, same task** setting of [2] and end with the extreme case of **different domain, different task**.
- We test using a variety of different attack using the TextAttack framework [3], which offers a variety of algorithms to craft adversarial examples in NLP.
- We used samples from the target domain to give the target model the best chance at defending against the attack. So if it fails here it is also likely to fail when we sample from the attack distribution.

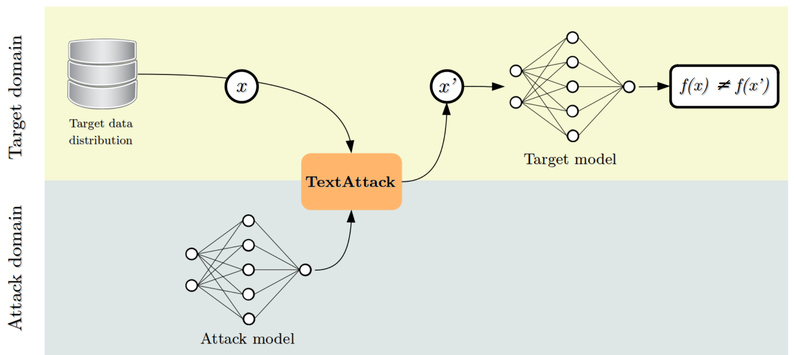


Figure 1. Illustration of the attack pipeline

## Test Procedure

1. Sample inputs from the target's data distribution (test set).
2. Using TextAttack, generate adversarial examples that successfully fool the attack model.
3. Finally, calculate the accuracy of the target model on those generated examples.

## Results

| Setting | Attack Model | Target Model |
|---|---|---|
| Similar domain, same task | roberta-base-imdb | roberta-base-rotten-tomatoes |
| Similar domain, different task | twitter-roberta-base-irony | twitter-roberta-base-sentiment |
| Different domain, same task | twitter-roberta-base-sentiment | roberta-base-rotten-tomatoes |
| Different domain, different task | twitter-roberta-base-irony | roberta-base-rotten-tomatoes |

Table 1. Summary of settings and models used

| Setting | Original accuracy | Attack | Relative decrease in accuracy |
|---|---|---|---|
| Similar domain, same task | 88.30 | BAEGarg2019 | 29.56 |
| | | DeepWordBugGao2018 | **30.80** |
| | | TextFoolerJin2019 | 24.01 |
| Similar domain, different task | 70.50 | BAEGarg2019 | **7.52** |
| | | DeepWordBugGao2018 | 4.82 |
| | | TextFoolerJin2019 | 5.39 |
| Different domain, same task | 88.30 | BAEGarg2019 | **17.44** |
| | | DeepWordBugGao2018 | 12.34 |
| | | TextFoolerJin2019 | 9.74 |
| Different domain, different task | 88.30 | BAEGarg2019 | 6.91 |
| | | DeepWordBugGao2018 | **8.61** |
| | | TextFoolerJin2019 | 6.91 |

Table 2. Target model accuracies in each setting (ordered by transferability), in bold is the biggest decrease of the three attacks.

| Setting | Original accuracy | Attack | Relative decrease in accuracy |
|---|---|---|---|
| Similar domain, same task | 95.00 | BAEGarg2019 | **14.42** |
| | | DeepWordBugGao2018 | 13.79 |
| | | TextFoolerJin2019 | 7.05 |
| Similar domain, different task | 73.46 | BAEGarg2019 | **6.07** |
| | | DeepWordBugGao2018 | 2.60 |
| | | TextFoolerJin2019 | 2.25 |
| Different domain, same task | 70.50 | BAEGarg2019 | **11.57** |
| | | DeepWordBugGao2018 | 10.63 |
| | | TextFoolerJin2019 | 7.32 |
| Different domain, different task | 84.69 | BAEGarg2019 | 4.52 |
| | | DeepWordBugGao2018 | **7.53** |
| | | TextFoolerJin2019 | 5.88 |

Table 3. Target model accuracies for **reversed settings** where the target model is treated as the attack model and vice versa for the attack model.

| | |
|---|---|
| Original sentence<br>Positive (76.92) | The closest thing to the experience of space travel |
| BAEGarg2019<br>Negative (89.04) | The closest thing to the notion of space travel |
| DeepWordBugGao2018<br>Negative (87.37) | The cloest thing to the sexperience of spacD Uravel |
| TextFoolerJin2019<br>Negative (96.45) | The nearest thing to the expertise of space travel |

Table 4. Examples of some of the attacks.

## Key Observations

- The further we are from the target model in terms of domain and task the less transferable the attack is (target model accuracy degradation isn't as large).
- Adversarial examples are more transferable across different domains than across different tasks, as indicated by a greater decrease in the **similar domain, different task** setting vs the **different domain, same task** setting.

## Conclusion

- For a **successful** adversary they must at least be aware of the domain or task, but attacks can be transferred even without this information.
- Having the same task is more important than having the same domain. We suspect this is due to how altering the task changes the behaviour of the last few layers of a model whereas altering the domain changes the first layers.
- The results give a clue to understand where adversarial examples in NLP come from. The high transferability in the **same task** settings indicate that they rely on higher-level features which suggests that adversarial examples in NLP are primarily concerned with the activations of the last layers.
- Further research is needed to understand exactly what about these last layers allows for their transferability.

## References

[1] M. Naseer, S. H. Khan, H. Khan, F. S. Khan, and F. Porikli, "Cross-domain transferability of adversarial perturbations," 2019.

[2] S. Datta, "Learn2weight: Parameter adaptation against similar-domain adversarial attacks," 2022.

[3] J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, "Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp," 2020.

[4] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, "Tweeteval: Unified benchmark and comparative evaluation for tweet classification," 2020.

[5] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," 2019.