Meng Thesis Transferability

Bipin Aasi

Feb 2023

Quick Summaries

0.1 Explainable AI: A Review of Machine Learning Interpretability Methods

Problem to Solve

- Survey paper

Contributions

- Liu et al. [131] while transferring non-targeted adversarial examples can be very effective in fooling neural networks, targeted adversarial examples do not work as well.
- A similar idea was later also proposed by Kuleshov et al. [154], which uses word replacement by greedy heuristics, while later Wang et al. [155] improved upon the genetic algorithm, achieving not only higher success rates, but also lower word substitution rates and more transferable adversarial examples when compared to [153].

0.2 Privacy and Security Issues in Deep Learning: A Survey [Liu et al., 2020b]

Problem to Solve

- Survey Paper

Contributions

- Gradients from Surrogate Models. The adversarial samples are transferability. That is, the gradient of the surrogate model is also helpful for attacking the target model. Therefore, they used the gradient of the surrogate model to guide the update direction of the boundary attack, which improves the attack efficiency
- To a certain extent, the above improvements improve the efficiency of the algorithm. However, the gradient of the surrogate model relies on the transferability of the model. Later, Chen et al. [141] further improved the boundary attack by utilizing Monte Carlo estimation to determine the direction of the gradient, which does not rely on the transferability of the model
- Ilyas et al. [218] pointed out that models trained on the same dataset tend to learn similar non-robust features, which accounts for the transferability of adversarial examples. However, why DL model tend to learn non-robust features and how to make them learn robust features is still an open question

Open problems:

- Perceivability: Need to propose methods that make the perturbations not only unperceivable, but preserve correct grammar and semantics

Personal Limitations:

- Only on images not language

0.3 Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey [Zhang et al., 2020]

Problem to Solve

- Survey of Adversarial attacks in NLP as of April 2019

Contributions

- Perturbation Measurement for text:
  - Grammar and syntax checker, Perplexity, Paraphrase
  - Semantic-preserving measurement: semantic similarity/distance is often performed on word vectors by adopting vectors' similarity/distance measurements (euclidean distance or Cosine Similarity)
  - Edit-based measurement, minimum changes from one string to the other: Levenshtein Distance, Word Mover's Distance, Number of changes, Jaccard similarity coefficient
- Summary of attacks (white box, granularity, etc) and their perturb control.
- Benchmark datasets for different langauge problems (classification, translation, summary, comprehension etc.)

Open problems:

- Perceivability: Need to propose methods that make the perturbations not only unperceivable, but preserve correct grammar and semantics
- Transferability: Organized into three levels in deep neural networks:
  - same architecture with different data
  - different architectures with same application
  - different architectures with different data

0.4 Adversarial Examples are not Bugs, they are Features [?]

Problem to Solve

- Claims paradigm that adversarial robustness is a goal, and can be disentangled and pursued independently from maximizing accuracy is wrong.
- Correct paradigm is: adversarial vulnerability is a direct result of sensitivity to well-generalizing features in the data

Contributions

- claim One of the most intriguing properties of adversarial examples is that they transfer across models with different architectures and independently sampled training sets
- Show that non-robust features alone are sufficient for good generalization
- The key differentiating aspect of our model is that adversarial perturbations arise as well-generalizing, yet brittle, features, rather than statistical anomalies

0.5 Leveraging transferability and improved beam search in textual adversarial attacks [Zhu et al., 2022]

Problem to Solve

- Prior works utilize different word replacement strategies to generate semantic-preserving adversarial texts. These query-based methods, however, have limited exploration of the search space

Contributions

- transferable vulnerability from surrogate models to choose vulnerable candidate words for target models. We empirically show that beam search with multiple random attacking positions works better than the commonly used greedy search with word importance ranking
- improved beam search which can achieve a higher success rate than the greedy approach under the same query budget

Personal Limitations

- Basis of transferability from surrogate models to perform textual adversarial attacks under the black-box setting for language models is wrong!

0.6 DELVING INTO TRANSFERABLE ADVERSARIAL EXAMPLES AND BLACK-BOX ATTACKS [?]

Problem to Solve

- Previous works mostly study the transferability using small scale dataset

Contributions

- Study both non-targeted and targeted adversarial examples, and show that while transferable non-targeted adversarial examples are easy to find, targeted adversarial examples generated using existing approaches almost never transfer with their target labels
- Propose novel ensemble-based approaches to generating transferable adversarial examples

Personal Limitations

- Done on images, not language

0.7 On the Transferability of Adversarial Attacks against Neural Text Classifier [Yuan et al., 2020]

Problem to Solve

- Systematically investigate transferability of adversarial examples for text classification and factors which influence transferability.

Contributions

- Investigated four critical factors of NLP neural models, including network architectures, tokenization schemes, embedding types, and model capacities and how they impact the transferability of text adversarial examples with more than sixty different models.
- Described a algorithm to discover highly-transferable adversarial word replacement rules that can be applied to craft adversarial examples with strong transferability across various neural models without access to any of them
- Compares the transferability rates. Required a base transferability rate
- Since those adversarial examples are modelagnostic, they provide an analysis of global model behavior and help to identify dataset biases.

Personal Limitations

- Focuses only on sentiment classification

0.8 Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment [Jin et al., 2020]

Problem to Solve

- A baseline to generate adversarial text for text classification and textual entailment.

Contributions

- effective—it outperforms previous attacks by success rate and perturbation rate
- utility-preserving—it preserves semantic content, grammaticality, and correct types classified by humans
- efficient—it generates adversarial text with computational complexity linear to the text length.
- Compares the transferability rates. Required a base transferability rate
- there is a moderate degree of transferability between models, and the transferability is higher in the textual entailment task than in the text classification task. Moreover, the adversarial samples generated based on the model with higher prediction accuracy, i.e. the BERT model here, show higher transferability

Personal Limitations

- Focuses only on sentiment classification and BERT

0.9 Universal Adversarial Triggers for Attacking and Analyzing NLP [Wallace et al., 2019]

Problem to Solve

- Creating a universal attack for language models which are transferable across models

Contributions

- Proposes a gradient-guided search over tokens which finds short trigger sequences (e.g., one word for classification and four words for language modeling) that successfully trigger the target prediction.
- Triggers are optimized using white-box access to a specific model, they transfer to other models for all tasks we consider
- Attacks 3 modes of language tasks: Text Classification, Question and Answering (Reading Comprehension), Text Generation GPT2 model
- For SQUAD: Uses baseline model and test the trigger's transferability to more advanced models (with different embeddings, tokenizations, and architectures).
- The baseline is BiDAF (Seo et al., 2017).
- Test the trigger's transferability to black-box models: QANet (Yu et al., 2018), an ELMo-based BiDAF model (ELMo), and a BiDAF model that uses character level convolutions (Char).

Future Work

- In future work, we aim to both attribute and defend against errors caused by adversarial triggers.

Personal Limitations

- Attacks are not adversarial because they are detectable by the human eye. Do not make semantic sense to append triggers such as "zoning tapping fiennes Visually imaginative, thematically instructive and thoroughly delightful, it takes us on a roller-coaster rid"

0.10 MALCOM: Generating Malicious Comments to Attack Neural Fake News Detection Models [Le et al., 2020]

Problem to Solve

- Attack SOTA fake news detectors

Contributions

- proposing a novel attack scenario against fake news detectors, in which adversaries can post malicious comments toward news articles to mislead SOTA fake news detectors
- Phrase I: identifying target articles to attack. Phrase II: generating malicious comments. Phrase III: appending generated comments on the target articles
- Developed an end-to-end adversarial comment generation framework to achieve such an attack (Malcom)
- We also compare our attack model with four baselines across two real-world datasets
- benchmark coherency We derive a topic coherency score of a set of arbitrary comments C and its respective set of articles X of size N as follows: Tk(X, C) = 1 N N i=0[1 cos(LDAk(xcontent i), LDAk(ci))], where cos(·) is a cosine similarity function. LDAk(·) is a Latent Dirichlet Allocation (LDA) model that returns the distribution of k different topics of a piece of text.
- quality and diversity: We use BLEU and negative-loglikelihood loss (NLL gen) scores to evaluate how well generated comments are in terms of both quality and diversity

Future Work

- Whether or not comments generated using one sub-domain (e.g., political fake news) can be transferable to another (e.g., health fake news) is also out of scope of this paper. Hence, we leave the investigation on the proposed attack's transferability across different datasets for future work. Moreover, we also plan to extend our method to attack graph based fake news detectors (e.g., [24]), and evaluate our model with other defense mechanisms such as adversarial learning

Personal Limitations

- Future work only considers transferability across datasets not models

0.11 T3: Tree-Autoencoder Constrained Adversarial Text Generation for Targeted Attack [Wang et al., 2019a]

Problem to Solve

- Build a target-controllable adversarial attack framework T3 for text (white box attack) for sentiment analysis and question answering (QA)

Contributions

- propose a tree-based autoencoder to embed the discrete text data into a continuous representation space, upon which we optimize the adversarial perturbation
- show that the generated adversarial texts have high transferability which enables the black-box attacks in practice
- adversarial examples have better transferability among the models with similar architectures than different architectures
- perform transferability based blackbox attacks. Specifically, the transferability-based blackbox attack uses adversarial text generated from whitebox BERT model to attack blackbox SAM, and vice versa.
- applied to regularize the syntactic correctness of the generated text and manipulate it on either sentence (T3(Sent)) or word (T3(Word)) level

Future Work

- adverserial training
- Interval Bound Propagation (IBP) (Dvijotham et al., 2018) is proposed as a new technique to theoretically consider the worst-case perturbation. with other defense mechanisms such as adversarial learning
- Language models including GPT2 (Radford et al., 2019) may also function as an anomaly detector

Personal Limitations

- Attacking very basic models (BERT and Self-Attentive Mode) using basic models for QA: Bi-Directional Attention Flow (BIDAF))

Problem to Solve

- Propose a fast gradient-based text adversarial attack , which is about 20 times faster than existing text attack methods and could achieve similar attack performance
- creates a defence that blocks transferability of adversarial attacks

Contributions

- Fast Gradient Projection Method (FGPM) based on synonym substitution, which is about 20 times faster than existing text attack methods and could achieve similar attack performance
- incorporate FGPM with adversarial training and propose a text defense method called Adversarial Training with FGPM enhanced by Logit pairing (ATFL). Experiments show that ATFL could significantly improve the model robustness and block the transferability of adversarial examples

0.12 Adversarial Training for Large Neural Language Models [Liu et al., 2020a]

Problem to Solve

- BERT are still vulnerable to adversarial attacks even after adversarial fine tuning.

Contributions

- Show that adversarial pretraining can improve both generalization and robustness
- Propose a general algorithm ALUM (Adversarial training for large neural LangUage Models), which regularizes the training objective by applying perturbations in the embedding space that maximizes the adversarial loss.

Open problems:

- further study on the role of adversarial pre-training in improving generalization and robustness; speed up adversarial training; apply ALUM to other domains

0.13 Model Extraction and Adversarial Transferability, Your BERT is Vulnerable! [?]

Problem to Solve

- Adversary can steal a BERT-based API service (the victim/target model) on multiple benchmark datasets with limited prior knowledge and queries

Contributions

- show that the extracted model can lead to highly transferable adversarial attacks against the victim model. even when there is an architectural mismatch between the victim model and the attack model
- Find that unless the performance of the victim model is sacrificed, both model extraction and adversarial transferability can effectively compromise the target models
- In phase 1, Model Extraction Attack (MEA) labels queries using the victim API, and then trains an extracted model on the resulting data. In phase 2, Adversarial Example Transfer (AET) generates adversarial typo examples on the extracted model, and transfers them to the victim API

Open problems:

- In the future, we plan to extend our work to more complex NLP tasks, and develop more effective defences

0.14 Are Transformers More Robust Than CNNs?[?]

Problem to Solve

- Argue that Transformers are much more robust than Convolutions Neural Networks o provide the first fair in-depth comparisons between Transformers and CNNs, focusing on robustness evaluations.

Contributions

- CNNs can easily be as robust as Transformers on defending against adversarial attacks, if they properly adopt Transformers' training recipes
- Ablations suggest such stronger generalization is largely benefited by the Transformer's self-attention-like architectures per se, rather than by other training setups.
- Generalization for Out of Distribution samples is not soley reliant on pretraining on (external) large-scale datasets

Personal Limitations

- Done on images, not Language

0.15 Pretrained Transformers Improve Out-of-Distribution Robustness[?]

Problem to Solve

- BERT achieve high accuracy on indistribution examples, do they generalize to new distributions? (In context of NLP Sentiment Analysis)

Contributions

- measure the generalization of previous models including bag-of-words models, ConvNets, and LSTMs, and we show that pretrained Transformers' performance declines are substantially smaller
- Ablations suggest such stronger generalization is largely benefited by the Transformer's self-attention-like architectures per se, rather than by other training setups.
- finding that larger models are not necessarily more robust
- distillation can be harmful
- more diverse pretraining data can enhance robustness

Open problems:

- while pretrained Transformers are moderately robust, there remains room for future research on robustness

NLP Attacks Tools

- https://github.com/QData/TextAttack
- https://github.com/robustness-gym/robustness-gym

References

[Bai et al., 2021] Bai, Y., Mei, J., Yuille, A. L., and Xie, C. (2021). Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, 34:26831–26843.

[Dai et al., 2019] Dai, J., Chen, C., and Li, Y. (2019). A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878.

[Du et al., 2020] Du, C., Sun, H., Wang, J., Qi, Q., and Liao, J. (2020). Adversarial and domain-aware bert for cross-domain sentiment analysis. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 4019–4028.

[Eger et al., 2019] Eger, S., Şahin, G. G., Rücklé, A., Lee, J.-U., Schulz, C., Mesgar, M., Swarnkar, K., Simpson, E., and Gurevych, I. (2019). Text processing like humans do: Visually attacking and shielding nlp systems. *arXiv preprint arXiv:1903.11508*.

[Gan et al., 2020] Gan, Z., Chen, Y.-C., Li, L., Zhu, C., Cheng, Y., and Liu, J. (2020). Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628.

[Goel et al., 2021] Goel, K., Rajani, N., Vig, J., Tan, S., Wu, J., Zheng, S., Xiong, C., Bansal, M., and Ré, C. (2021). Robustness gym: Unifying the nlp evaluation landscape. *arXiv preprint arXiv:2101.04840*.

[Gunel et al., 2020] Gunel, B., Du, J., Conneau, A., and Stoyanov, V. (2020). Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

[Hambardzumyan et al., 2021] Hambardzumyan, K., Khachatrian, H., and May, J. (2021). Warp: Word-level adversarial reprogramming. *arXiv preprint arXiv:2101.00121*.

[He et al., 2021] He, X., Lyu, L., Xu, Q., and Sun, L. (2021). Model extraction and adversarial transferability, your bert is vulnerable! *arXiv preprint arXiv:2103.10013*.

[Hendrycks et al., 2020] Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. (2020). Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*.

[Jiang et al., 2019] Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Zhao, T. (2019). Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*.

[Jin et al., 2020] Jin, D., Jin, Z., Zhou, J. T., and Szolovits, P. (2020). Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

[Karimi et al., 2021] Karimi, A., Rossi, L., and Prati, A. (2021). Adversarial training for aspect-based sentiment analysis with bert. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8797–8803. IEEE.

[Le et al., 2020] Le, T., Wang, S., and Lee, D. (2020). Malcom: Generating malicious comments to attack neural fake news detection models. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 282–291. IEEE.

[Li et al., 2020] Li, L., Ma, R., Guo, Q., Xue, X., and Qiu, X. (2020). Bert-attack: Adversarial attack against bert using bert. *arXiv preprint arXiv:2004.09984*.

[Li et al., 2021] Li, S., Liu, H., Dong, T., Zhao, B. Z. H., Xue, M., Zhu, H., and Lu, J. (2021). Hidden backdoors in human-centric language models. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 3123–3140.

[Liu et al., 2020a] Liu, X., Cheng, H., He, P., Chen, W., Wang, Y., Poon, H., and Gao, J. (2020a). Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*.

[Liu et al., 2020b] Liu, X., Xie, L., Wang, Y., Zou, J., Xiong, J., Ying, Z., and Vasilakos, A. V. (2020b). Privacy and security issues in deep learning: A survey. *IEEE Access*, 9:4566–4593.

[Morris et al., 2020] Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., and Qi, Y. (2020). Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909*.

[Perez and Ribeiro, 2022] Perez, F. and Ribeiro, I. (2022). Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*.

[Qiu et al., 2019] Qiu, S., Liu, Q., Zhou, S., and Wu, C. (2019). Review of artificial intelligence adversarial attack and defense technologies. *Applied Sciences*, 9(5):909.

[Ren et al., 2019] Ren, S., Deng, Y., He, K., and Che, W. (2019). Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1085–1097.

[Wallace et al., 2019] Wallace, E., Feng, S., Kandpal, N., Gardner, M., and Singh, S. (2019). Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*.

[Wang et al., 2019a] Wang, B., Pei, H., Pan, B., Chen, Q., Wang, S., and Li, B. (2019a). T3: Tree-autoencoder constrained adversarial text generation for targeted attack. *arXiv preprint arXiv:1912.10375*.

[Wang et al., 2019b] Wang, D., Gong, C., and Liu, Q. (2019b). Improving neural language modeling via adversarial training. In *International Conference on Machine Learning*, pages 6555–6565. PMLR.

[Wang et al., 2020] Wang, T., Wang, X., Qin, Y., Packer, B., Li, K., Chen, J., Beutel, A., and Chi, E. (2020). Cat-gen: Improving robustness in nlp models via controlled adversarial text generation. *arXiv preprint arXiv:2010.02338*.

[Wang et al., 2019c] Wang, X., Jin, H., and He, K. (2019c). Natural language adversarial attack and defense in word level.

[Wang et al., 2021] Wang, X., Yang, Y., Deng, Y., and He, K. (2021). Adversarial training with fast gradient projection method against synonym substitution based text attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13997–14005.

[Yuan et al., 2020] Yuan, L., Zheng, X., Zhou, Y., Hsieh, C.-J., and Chang, K.-W. (2020). On the transferability of adversarial attacksagainst neural text classifier. *arXiv preprint arXiv:2011.08558*.

[Zhang et al., 2020] Zhang, W. E., Sheng, Q. Z., Alhazmi, A., and Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.

[Zhu et al., 2022] Zhu, B., Gu, Z., Qian, Y., Lau, F., and Tian, Z. (2022). Leveraging transferability and improved beam search in textual adversarial attacks. *Neurocomputing*, 500:135–142.

[Zhu et al., 2019] Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T., and Liu, J. (2019). Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*.