GLARE: Generative Left-to-right AdversaRial Examples

Stanford CS224N Custom Project

Ryan A. Chi
Department of Computer Science
Stanford University
ryanchi@cs.stanford.edu

Nathan Kim
Department of Computer Science
Stanford University
nathangk@stanford.edu

Zander Lack
Department of Computer Science
Stanford University
zanderll@stanford.edu

Abstract

Recently, transformer models [1] have been applied to adversarial example generation—word level substitution models utilizing BERT [2] ([3], [4], [5]) have out-performed previous state-of-the-art approaches. Extending the paradigm of transformer-based generation of adversarial examples, we propose a novel textual adversarial example generation framework based on transformer language models: our method (GLARE) generates word- and span-level perturbations of input examples using ILM [6], a GPT-2 language model finetuned to fill in masked spans. We demonstrate that GLARE achieves a superior performance to CLARE (the current state-of-the-art model) in terms of attack success rate and semantic similarity between the perturbed and original examples.

1 Key Information to include

- CS 224N Mentor: Shikhar Murty
- Stanford AI Mentor: Ethan A. Chi
- External Collaborators: N/A
- Sharing Project: N/A

2 Introduction

Adversarial examples, though well-studied in the computer vision domain, are difficult to produce in the natural language processing domain, partly due to the discrete nature of text. Previous approaches, drawing on constrained methods such as rule-based heuristics [7] and synonym substitution [8], have attained relatively limited success, largely because these approaches consider neither syntactic nor semantic structure. As a result, adversarial examples yielded by these models often suffer from a lack of grammaticality, idiomaticity, and overall fluency.

These shortcomings can be explained in terms of model size and complexity. In theory, it is possible for a model of any size to generate adversarial output; however, largely due to the discrete nature of text, it is far likelier to learn an accurate representation of the search space using a highly parameterized model such as a transformer. Previous transformer-based textual adversarial frameworks have used BERT to generate word-level replacements, essentially re-purposing its pretext task (masked token prediction). Yet BERT's training objective is not text generation—this is, however,

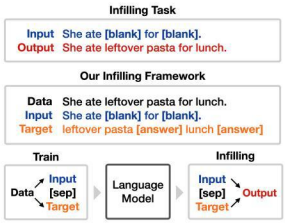

Figure 1: A diagram illustrating an example of a sentence that was selectively masked and infilled by GPT-2 according to the ILM framework developed by Donahue et al.

precisely the explicit objective of another class of models: generative language models, often simply referred to as *language models*. We predict that a generative adversarial example framework utilizing GPT-2 will at least match (and perhaps outperform) CLARE, as well as the similar BERT-based frameworks BAE [3] and BERT-ATTACK [5]. Furthermore, we predict that a framework based on GPT-2 (trained on 8 million webpages) will outperform one trained on BERT (trained only on BooksCorpus and Wikipedia) on the task of generalizing to previously unseen text.

However, left-to-right LMs such as GPT-2 suffer from the weakness of not being able to incorporate context from both sides of the current position in the sentence. To allow the LM access to both the left and right context at a given point in the sentence, we adopt the strategy introduced by (and adapt code written by) Donahue et al. [6]: the words to infill are replaced by special `[blank]` tokens, and for a given sentence, the model learns to continue the input with the sequence `[answer]` $\texttt{word}_1$ `[answer]` $\texttt{word}_2$ `...` $\texttt{word}_n$ `[end_infill]`, where each word or sequence of words preceding the nth `[answer]` token corresponds to model's suggested replacement for the nth masked word. Although infilling with GPT-2 is not novel, applying the method to better attack a masked language model has not been previously done.

3 Related Work

- **TEXTFOOLER [9]** is a lightweight method that utilizes word importance ranking and synonym substitution (as well as POS constancy and semantic similarity constraints) to create adversarial examples.
- **BERTATTACK [5]** is a BERT-based method that uses word-importance ranking to replace individual tokens in descending order of importance.
- **BAE [3]** is highly similar to BERTATTACK with the additional ability to insert tokens at positions neighboring a replaced token.
- **CLARE [4]** is similar to its concurrent papers **BAE** and **BERTATTACK** with several small improvements: the authors substitute RoBERTa [10] for BERT and draws from three different perturbations (*replace*, *insert*, and *merge*) rather than simply one or two.

We consider our work to be a natural progression of BERTATTACK, BAE, and CLARE, as we substitute GPT-2 for the BERT-based model (with the expectation of generating higher-quality infills) and further allow for variable-length merges and insertions (capped by a user-specified length).

4 Approach

Similar to the CLARE paper, our method generates adversarial examples for untargeted attacks on sequence classification tasks. We create an attack recipe which, given a victim model and a dataset, creates new test examples from examples in the dataset which retain a large degree of similarity with the source example yet are classified differently by the victim model. To carry out this task, we use TEXTATTACK [11], a Python framework for producing adversarial attacks derived from a wide selection of approaches and carried out on a variety of models. Specifically, the custom attack recipe consists of word-level replacements supplied by a fine-tuned version of an infilling GPT-2 model and constrained by the minimum sentence-wise cosine similarity score in a given example. Our baselines

consist of the current SOTA model (CLARE) and its concurrent model BERTATTACK as well as the former state-of-the-art model TEXTFOOLER. [1]

5 Experiments

5.1 Experimental Settings

In addition to our baselines, we evaluate on three variants of the GLARE model. **GLARE$_{\text{single}}$** is the simplest variant, and replaces single words with single words using a fine-tuned ILM model in an iterative fashion until the attack succeeds or no more words can be replaced. Following [9], words are ordered in descending order of their Word Importance Ranking (WIR), a measure of the effect on a given example's classification after the word in question is replaced with an `UNK` token. **GLARE$_{\text{multi}}$** leverages the ILM model's ability to fill in short n-grams, and replaces multi-word spans in the input with spans of variable length. Words are again ordered in descending WIR scores; however, contiguous subsequences in the ranked wordlist which also constitute contiguous subarrays of the original example (in any order) are collapsed into single spans which are replaced in tandem by the model. **GLARE$_{\text{multi}}$** introduces two additional parameters: a maximum extension e_{max}, which limits the absolute difference between the length of an input span and its replacement, and a maximum collapse width c_{max}, which sets an upper bound on the number of words in the ranked wordlist which can be collapsed into a sequence. In all our experiments, we use $e_{\text{max}} = 5$ and $c_{\text{max}} = 3$. Finally, we evaluate on **GLARE$_{\text{OOD}}$**, an out-of-domain variant of **GLARE$_{\text{single}}$** which uses an ILM model trained on the ROCStories short story corpus [12] (as provided by the authors) in place of the fine-tuned dataset.

DATASET	TRAIN	TEST	SUPPORT
Yelp Polarity	560K	38K	{0 (negative), 1 (positive)}
AG News	120K	7.6K	{0 (world), 1 (sports), 2 (business), 3 (sci/tech)}
MultiNLI	393K	20K	{0 (entailment), 1 (neutral), 2 (contradiction)}
QNLI	107K	5.5K	{0 (entailment), 1 (not entailment)}

Table 1: Dataset details

In accordance with previous textual adversarial methods [13], [4], we finetune the infilling GPT-2 model on the training data and evaluate on a test set of size 1000, each example with a length of no more than 100 words.

5.2 Evaluation method

We evaluate on the following metrics:

- **Attack Success Rate (A-rate)** is the percentage of attacks that were successfully performed.
- **Modification Rate (Mod)** is the average percentage of words in an example that were modified.
- **Perplexity (PPL)** as measured by a small (12-layer, 768-hidden, 12-heads, 117M parameters) non-finetuned GPT-2 model.
- **Grammar Error (GErr)** is the average number of grammatical errors introduced by each perturbed example.
- **Semantic Similarity (Sim)** is the cosine similarity between the original and perturbed text, as calculated by the Universal Sentence Encoder [14].
- **Average # of Queries** refers to the average number of substitutions it takes to produce each adversarial example in the dataset, giving a measure of GLARE's latency.

[1] Due to difficulties implementing the TEXTFOOLER and CLARE models with TEXTATTACK, the baseline values included in Table 2 were taken from [4].

	Yelp (PPL = 53.4)					**AG News** (PPL = 38.03)				
Model	**A-rate↑**	**Mod↓**	**PPL↓**	**GErr↓**	**Sim↑**	**A-rate↑**	**Mod↓**	**PPL↓**	**GErr↓**	**Sim↑**
GLARE (single-word)	77.0	16.6	163.3	1.23	0.70	56.1	23.3	331.3	1.43	0.69
GLARE (variable-len)	92.09	56.68	48.2	0.22	0.92	78.95	69.77	63.91	1.69	0.88
GLARE (single, OOD)	93.53	11.22	63.61	0.15	0.92	70.32	18.87	124.38	0.27	0.86
CLARE	79.7	10.3	51.5	83.5	0.25	79.1	6.1	62.8	86	0.17
0.76										
TEXTFOOLER	77.0	16.6	163.3	1.23	0.70	56.1	23.3	331.3	1.43	0.69
BERTATTACK	71.8	10.7	90.8	0.27	0.72	63.4	7.9	90.6	0.25	0.71
	MNLI (PPL = 28.87)					**QNLI** (PPL = 37.88)				
Model	**A-rate↑**	**Mod↓**	**PPL↓**	**GErr↓**	**Sim↑**	**A-rate↑**	**Mod↓**	**PPL↓**	**GErr↓**	**Sim↑**
GLARE (single-word)	92.92	6.16	77.94	0.23	0.84	86.89	10.04	72.92	0.22	0.87
GLARE (variable-len)	84.15	18.76	60.21	0.33	0.82	79.63	42.24	55.59	0.47	0.89
GLARE (single, OOD)	93.64	5.84	64.55	0.15	0.84	91.08	9.74	77.31	0.18	0.87
CLARE	88.1	7.5	82.7	0.02	0.82	83.8	11.8	76.7	0.01	0.78
TEXTFOOLER	59.8	13.8	161.5	0.63	0.73	57.8	16.9	164.6	0.62	0.72
BERTATTACK	82.7	8.4	86.7	0.04	0.77	76.7	13.3	86.5	0.03	0.73

Table 2: Adversarial example generation performance in attack success rate (A-rate), modification rate (Mod), perplexity (PPL), number of increased grammar errors (GErr), and textual similarity (Sim). The perplexity of each dataset is marked in the header. ↑ (↓) represents which direction is more desirable. (Formatting is inspired by [4]).

5.3 Experimental details

All experiments were conducted on Microsoft Azure. We fine-tune our models using the same parameters outlined in [6], training each for at least 4000 steps. Further details can be found in Table 3 of the Appendix.

5.4 Results

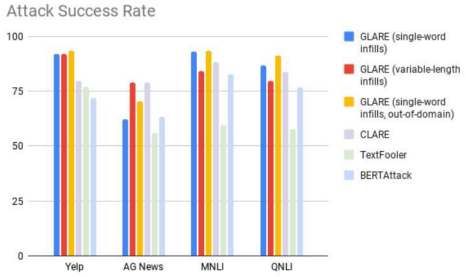

Figure 2: Comparison of attack success rates by different models.

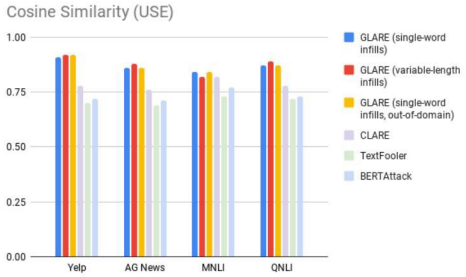

Figure 3: Comparison of cosine similarity scores between original and perturbed text by different models.

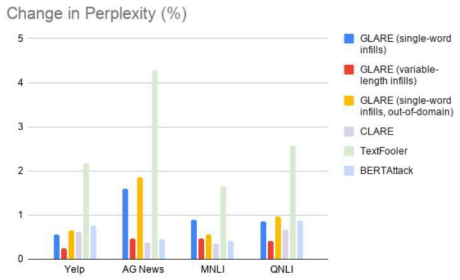

Figure 4: Comparison of increases in perplexity from original to perturbed text by different models.

6 User Study

In addition to quantitative metrics, we conducted a user study on the AG News data set to assess the more qualitative aspects of our results. Our technique was modeled off the study done in [4], albeit at a smaller scale. To compare the outputs of GLARE (single-word infills) and GLARE (variable-length infills), we presented the user with an original text, and a successful adversarial attack from both model variants (although the user is unaware which perturbed text corresponds to which model variant). The user then had to decide which perturbed text was closer in meaning to the the original, and which sentence was more grammatically and syntactically correct. In addition, we also asked the user to classify successfully attacked examples from each model variant as one of the AG News classes.

During our small user study (50 examples), we found that single-word infill variant of GLARE produced adversarial examples that were both more similar in meaning to the original text (59.18% of the time more similar than variable-length infills) and more grammatically and syntactically correct (61.22%). For both model variants, the class of the original text was correctly assessed 39.58% of the time.

	Single	**Multi**
% more correct than other method	**59.18%**	40.82%
% more grammatical than other method	**61.23%**	38.77%
% of the time same as original class	39.58%	39.58%

Table 3: Results of human evaluation. The first two rows compare the two methods against each other.

7 Analysis

Example Length: We note that longer inputs generally experience higher similarity scores when comparing their perturbed and original examples. We believe this is because the longer context gives the model a wider range of opportunities to perform an adversarial attack, as well as allowing the model a better glimpse into the semantic and syntactic structure of the example. Our qualitative analysis corroborates this hypothesis—the authors find that variable-length infills substantially improve the fluency of adversarial examples drawn from longer inputs. However, this finding could potentially be simply a consequence of the fact that for longer examples, the percentage of changed words is bound to be lower for the same number of transformations and thus the sentences are more likely to be adjudged similar.

Modification Rate: We note that our model suffers from a higher modification rate than CLARE. Although this is ostensibly undesirable, one benefit of a larger modification rate is that attacks are less likely to comprise simple polarity switches (e.g., "The food was delicious" $\rightarrow$ "The food was terrible"), which feature low modification rates but are not satisfactory adversarial examples as they

necessitate a change int the example's gold label. That said, we would still like to strive for a lower modification rate while maintaining the same fluent adversarial substitutions.

AG News: Of our four datasets, the model attacks AG News least successfully. We observe this, noting that the single-word-infill non-finetuned GLARE model performs at or above the level of the finetuned GLARE model. This could be explained by multiple causes: perhaps the title case of the AG news dataset causes training to be more difficullt. Alternatively, the broader domain results in a higher likelihood of encountering tokens that are outside of the model's training distribution.

8 Conclusion

We are able to successfully outputerform CLARE (the current SOTA) on a number of metrics: specifically, attack success rate, perplexity, and semantic similarity. The primary limitations of our work are that fine-tuning the GPT-2 model does not appear to as successful as we would like, demonstrated by the fact that the ILM model finetuned on stories was able to often match or even outperform the corresponding model finetuned on the specific dataset.

It is debatable whether generating adversarial examples is an end in itself, but they indirectly lead language models to learn better representations by allowing them to defend against adversarial attacks, a conclusion that the CLARE authors as well as others come to [9]. Our next step is to compare the performance of pre-trained models with and without adversarial fine-tuning to determine if they truly increase a model's robustness.

9 Acknowledgments

We would like to thank our Stanford NLP mentor Ethan A. Chi for his comprehensive and insightful advice throughout the course of this project. Furthermore, we are grateful to our CS 224N mentor Shikhar Murty for his thoughtful and encouraging suggestions regarding our proposal and milestone. Additionally, we would like to recognize Microsoft Azure for sponsoring this project. Finally, we would like to thank Prof. Chris Manning, John Hewitt, and all of the TAs and coordinators for a great quarter filled with engaging, illuminating lectures and assignments that were thought-provoking, challenging, and—above all—as empowering as they were wonderfully satisfying.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Siddhant Garg and Goutham Ramakrishnan. Bae: Bert-based adversarial examples for text classification, 2020.

[4] Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. Contextualized perturbation for textual adversarial attack. 2020.

[5] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. Bert-attack: Adversarial attack against bert using bert, 2020.

[6] Chris Donahue, Mina Lee, and Percy Liang. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[7] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. Deep text classification can be fooled, Jan 2019.

[8] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples, Jan 2018.

[9] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? natural language attack on text classification and entailment. *arXiv preprint arXiv:1907.11932*, 2019.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[11] John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. 2020.

[12] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California, June 2016. Association for Computational Linguistics.

[13] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani B. Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. *CoRR*, abs/1804.07998, 2018.

[14] Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder, 2018.

A Appendix

A.1 Training Details

DATASET	TRAINING STEPS
Yelp Polarity	8740
AG News	8383
MultiNLI	18995
QNLI	4113

Table 4: Dataset details

A.2 Example Output

```
423
Great pub - good food.
Not a pub - good food.
```

Figure 5: Example 1.

A.3 Illustration of GLARE Model

```
42
Very high quality food. Friendly service and a clean environment.
Very high expectations for the food. Great service and a clean environment.
```

Figure 6: Example 2.

Figure 7: An artist's representation of the GLARE model.