INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

ADVERSARIAL NLI

ASHNA DUA (2021101072)
ASHNA.DUA@STUDENTS.IIIT.AC.IN
VANSHITA MAHAJAN (2021101102)
VANSHITA.MAHAJAN@STUDENTS.IIIT.AC.IN
PRISHA (2021101075)
PRISHA.KUMAR@RESEARCH.IIIT.AC.IN

SEPTEMBER 18, 2024

Contents

1

Problem Statement

1.1 Introduction

Large-scale pre-trained language models like BERT and RoBERTa have significantly advanced performance in numerous Natural Language Processing (NLP) tasks, establishing themselves as essential tools in the field. Despite their remarkable capabilities, these models exhibit vulnerabilities when exposed to adversarial textual inputs.

This creates a critical challenge, as even small perturbations in the input can lead to dramatic changes in model output, compromising their reliability in sensitive applications. Traditional fine-tuning approaches have fallen short in mitigating these weaknesses, leaving room for new, more resilient strategies.

In this project, we tackle the problem from an information-theoretic perspective. Our goal is to develop a robust fine-tuning framework that not only preserves the models' high performance but also increases their resistance to adversarial examples. By introducing mutual-information-based regularizers, we aim to reduce noise in the feature representations while reinforcing critical features, thus ensuring the model can maintain accuracy under both standard and adversarial training conditions.

1.2 Dataset: Adversarial Natural Language Inference (ANLI)

The Adversarial Natural Language Inference (ANLI) dataset is a large-scale, challenging benchmark designed for natural language understanding (NLU) tasks, specifically natural language inference (NLI). ANLI was collected using a unique iterative, adversarial human-and-model-in-the-loop process aimed at creating examples that expose model weaknesses, ensuring continuous learning and robustness.

1.2.1 Collection Process

The dataset was collected in three rounds (A1, A2, and A3), with increasing levels of difficulty in each round. The data collection followed the **Human-And-Model-in-the-Loop Enabled Training (HAMLET)** process, which involves human annotators attempting to create hypothesis examples (given a context and a target label) that state-of-the-art models misclassify. The process includes:

- **Context Selection:** Annotators were provided with short, multi-sentence passages (context) from sources such as Wikipedia, news articles, and other genres.
- **Hypothesis Generation:** Annotators wrote hypotheses corresponding to a target label (*entailment, contradiction, or neutral*) that aimed to fool the model into incorrect classifications.
- **Verification:** Hypotheses that fooled the model were then verified by human annotators to ensure that the label was correct and the example was valid.

The dataset includes both verified correct examples and those where the model failed, to continuously challenge the evolving model.

1.2.2 Rounds of Data Collection

Each round in ANLI represents a progressively more difficult phase of model training and evaluation:

- **A1 (Round 1):** The first round used contexts sampled from Wikipedia and the HotpotQA dataset. A BERT-Large model was employed as the baseline model for this round.
- **A2 (Round 2):** In the second round, a stronger RoBERTa model was used, incorporating the training data from Round 1 along with new data.
- **A3 (Round 3):** The third round involved a more diverse set of contexts from domains such as news articles, fiction, spoken text from court transcripts, and procedural texts, making it the most challenging.

The increasing difficulty in each round stems from the iterative nature of the data collection process, where models are strengthened and tested on progressively harder examples generated by human annotators.

To prevent overfitting to specific annotator styles or model characteristics, the test set for each round includes an **exclusive subset**. This exclusive subset consists of examples from annotators who did not contribute to the training data, ensuring that models are tested on unseen data from entirely new annotators. This prevents models from learning biases inherent to specific annotators and ensures better generalization across varied inputs.

1.2.3 Dataset Statistics

ANLI contains a total of 162,865 training examples collected over three rounds, with 3,200 examples in both the development and test sets for each round. The dataset is balanced to ensure no overfitting occurs due to model bias or specific genre characteristics. Table 1.1 provides a detailed breakdown of the dataset's statistics.

Round	Train Size	Dev Size	Test Size
A1	16,946	1,000	1,000
A2	45,460	1,000	1,000
A3	100,459	1,200	1,200

Table 1.1: *Dataset statistics for ANLI.*

1.2.4 Comparison to Other Datasets

ANLI represents an improvement over prior NLI datasets like SNLI and MNLI by focusing on more complex and longer contexts. While SNLI used image captions as contexts, ANLI sources multi-sentence passages from various genres, adding greater complexity. The human-and-model-in-the-loop process also makes ANLI more robust against overfitting, allowing models trained on this dataset to better handle real-world NLI tasks by addressing more nuanced and challenging examples.

2

Literature Review

Textual adversarial attacks have emerged as a significant concern for language models, particularly those based on pre-trained transformers like BERT and RoBERTa. The existing adversarial attacks predominantly focus on word-level manipulations. Ebrahimi et al., 2018 were among the first to introduce a white-box, gradient-based approach to search for adversarial word or character substitutions. Subsequent research (Alzantot et al., 2018; Ren et al., 2019; Zang et al., 2020; Jin et al., 2020) sought to refine these methods by restricting the perturbation search space and using Part-of-Speech (POS) checking to ensure that the adversarial examples maintain a natural appearance to human observers.

To defend against such adversarial attacks, three primary defense strategies have been proposed:

- **Adversarial Training**: A popular and practical defense mechanism, adversarial training augments the model with adversarial examples. Some works (C. Zhu et al., 2020; Jiang et al., 2020; Liu et al., 2020; Gan et al., 2020) employ PGD-based (Projected Gradient Descent) attacks in the embedding space to generate adversarial examples for data augmentation or use virtual adversarial training to regularize the objective function. However, the primary limitation of this approach is its dependency on the threat model, which makes it less effective against unseen attacks.
- **Interval Bound Propagation (IBP)**: Introduced by Dvijotham et al., 2018, IBP is designed to handle worst-case perturbations theoretically. Recent research (Huang et al., 2019; Jia et al., 2017) has extended this approach to the NLP domain, applying IBP to certify the robustness of language models. However, IBP relies on strong assumptions about model architecture, making it difficult to apply to modern transformer-based models.
- **Randomized Smoothing**: Cohen et al., 2019 introduced randomized smoothing as a method to guarantee robustness in l_2 norm by adding Gaussian noise to smooth the classifier. Ye et al., 2020 adapted this idea to the NLP domain by replacing Gaussian noise with synonym words, ensuring that adversarial word substitutions

within predefined synonym sets maintain robustness. However, ensuring the completeness of the synonym set presents a challenge in practice.

S. Zhu et al., 2020 addressed this issue by proposing a robust representation learning approach that considers the mutual-information perspective in the context of worst-case perturbation. However, their work primarily focused on continuous input spaces like those in computer vision. In contrast, InfoBERT adopts a novel information-theoretic perspective that is designed for both standard and adversarial training in the discrete input space typical of language models.

The existing literature highlights the progress made in defending against textual adversarial attacks and improving the robustness of language models through various techniques. However, most prior work has focused on either adversarial training or theoretical robustness certifications, both of which come with limitations in terms of model adaptability or threat model assumptions. InfoBERT takes a step forward by integrating mutual-information-based regularizers into the fine-tuning process, which not only enhances robustness in adversarial settings but also aligns with standard training requirements for discrete NLP input spaces. This combination of information theory and adversarial training positions InfoBERT as a unique contribution to the ongoing effort to make language models more robust and reliable.

3 Proposal

This project proposes the implementation of the **InfoBERT** training objective, followed by training and evaluation on the Adversarial Natural Language Inference (ANLI) dataset. The goal is to demonstrate the effectiveness of robust fine-tuning using InfoBERT as compared to traditional BERT-based models, particularly on the challenging ANLI dataset, which is designed to test natural language inference models with adversarially generated data.

3.1 Baseline Approach

The baseline for this project involves training a simple BERT model on the ANLI dataset. This BERT-based baseline will serve as the reference point for evaluating the performance improvements introduced by the robust training of InfoBERT. By establishing this baseline, we can assess the inherent difficulty of the ANLI dataset and the model's susceptibility to adversarial attacks.

3.2 InfoBERT Objective

InfoBERT introduces two mutual-information-based regularizers to enhance the robustness of pre-trained language models:

- **Information Bottleneck (IB) Regularizer:** This regularizer reduces the noisy mutual information between the input and feature representations, removing irrelevant details that may introduce adversarial vulnerability.
- **Anchored Feature Regularizer:** This regularizer enhances the mutual information between local stable features and global features, ensuring that useful and stable features are prioritized over noisy or adversarially modified ones.

The InfoBERT framework will be implemented and fine-tuned on the ANLI dataset to assess its performance against adversarially generated examples.

3.3 Evaluation

The performance of InfoBERT will be compared against the baseline BERT models trained on the ANLI dataset. Metrics such as accuracy, robust accuracy, and F1 scores will be reported to illustrate the gains in robustness achieved through the InfoBERT objective. Additionally, the results will highlight the challenging nature of the ANLI dataset by examining how both models perform under adversarial conditions.

3.4 Expected Outcome

The anticipated outcome is that InfoBERT will outperform the BERT-based baseline models on ANLI, particularly in terms of robust accuracy. The comparison will showcase the benefits of incorporating mutual-information-based regularization in handling adversarial data. This project will also demonstrate the difficulty of the ANLI dataset and its effectiveness in evaluating model robustness.

4

Methodology

The proposed methodology focuses on enhancing the robustness of language models against adversarial attacks by integrating two techniques: the **Information Bottleneck (IB) Regularizer** and the **Local Anchored Feature Regularizer**. Each technique addresses different aspects of model robustness and feature stability, providing a comprehensive approach to improving language model performance in adversarial settings.

4.1 Information Bottleneck (IB) Regularizer

The **IB Regularizer** is introduced to manage the trade-off between retaining useful information for accurate predictions and minimizing the complexity of the model's internal representations. The primary goal is to ensure that the intermediate representations learned by the model (denoted as T) are both ***informative*** for predicting the target label Y and ***compressed*** enough to avoid overfitting to noise or irrelevant details. This regularizer operates under the principle that a good representation should capture sufficient information about Y while ignoring unnecessary information about X (the input data).

4.1.1 Objective Function

The balance is achieved by formulating the optimization problem as follows:

$$L_{\text{IB}} = I(Y;T) - \beta I(X;T) \tag{4.1}$$

where:

- $I(Y;T)$ represents the mutual information between the target label Y and the representation T. This term aims to maximize the information that T contains about Y, ensuring that the representations are informative for predicting the target.
- $I(X;T)$ denotes the mutual information between the input data X and the representation T. This term penalizes excessive information about X captured by T, which helps to prevent overfitting and reduces the model complexity.

- $\beta > 0$ is a hyperparameter that controls the trade-off between the two terms. A larger β increases the emphasis on reducing the complexity of T, while a smaller β places more importance on retaining information about Y.

This formulation ensures that the learned representation T retains relevant information for prediction while ignoring unnecessary details from the input data, thus improving the generalization and robustness of the model.

4.2 Local Anchored Feature Regularizer

The goal of the local anchored feature extraction is to find features that carry useful and stable in- formation for downstream tasks. Instead of directly searching for local anchored features, we start with searching for nonrobust and unuseful features. To identify local nonrobust features, we perform adversarial attacks to detect which words are prone to changes under adversarial word substitution. We consider these vulnerable words as features nonrobust to adversarial threats. Therefore, global robust sentence representations should rely less on these vulnerable statistical clues.

During the local anchored feature extraction, we perform "virtual" adversarial attacks that generate adversarial perturbations in the embedding space, which abstracts the general idea behind word-level adversarial attacks.

4.2.1 Mathematical Representation

Formally, given an input sentence $x = [x_1; x_2; \ldots; x_n]$ with its corresponding local embedding representation $t = [t_1; \ldots; t_n]$, where x and t are realizations of random variables X and T, we generate an adversarial perturbation δ in the embedding space to increase the task loss $\mathcal{L}_{\text{task}}$.

The adversarial perturbation δ is initialized to zero, and the gradient of the loss with respect to δ is calculated as:

$$g(\delta) = \nabla_\delta \mathcal{L}_{\text{task}}(q_\psi(t + \delta), y)$$

The perturbation is updated as:

$$\delta \leftarrow \boldsymbol{Q}_{\|\delta\|_F \leq \epsilon} \left(\frac{\eta g(\delta)}{\|g(\delta)\|_F} \right)$$

A feasible plan is to choose the words whose perturbation is neither too large (nonrobust features) nor too small (un-useful features), e.g., the words whose perturbation rankings are among ***50% to 80%*** of all the words.

4.3 Final Objective Function

By incorporating the term $I(T_i; Z)$ into the previous objective function, the final objective function becomes:

$$\max I(Y;T) - n\beta_X \sum_{i=1}^{n} I(X_i;T_i) + \alpha_X \sum_{i=1}^{M} I(T_{kj};Z)$$

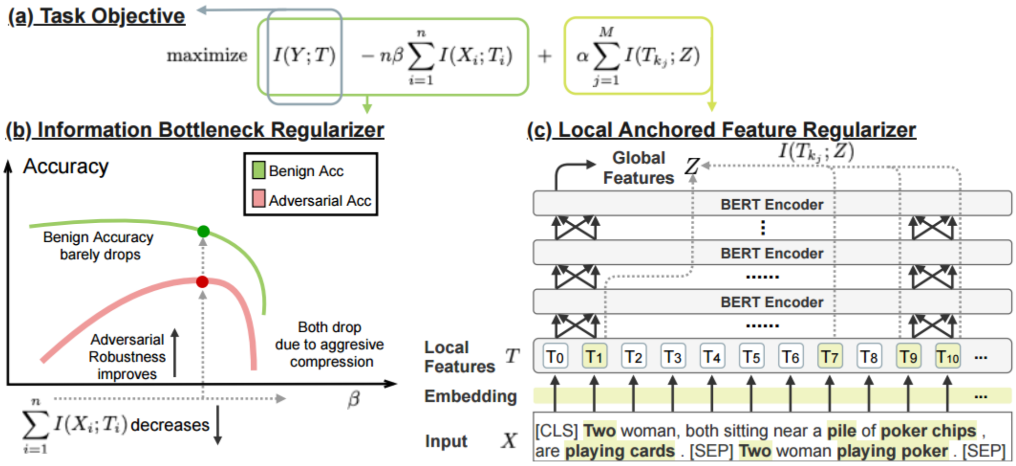
(a) Task Objective
maximize $I(Y;T)$ $-n\beta\sum_{i=1}^{n} I(X_i;T_i)$ + $\alpha\sum_{j=1}^{M} I(T_{k_j};Z)$
(b) Information Bottleneck Regularizer
Accuracy
🟩 Benign Acc
🟥 Adversarial Acc
Benign Accuracy barely drops
Adversarial Robustness improves
Both drop due to aggresive compression
β
$\sum_{i=1}^{n} I(X_i;T_i)$ decreases
(c) Local Anchored Feature Regularizer
Global Features Z
$I(T_{k_j};Z)$
BERT Encoder
BERT Encoder
BERT Encoder
Local Features T
T0 T1 T2 T3 T4 T5 T6 T7 T8 T9 T10 ...
Embedding
Input X
[CLS] Two woman, both sitting near a pile of poker chips , are playing cards . [SEP] Two woman playing poker . [SEP]

Figure 4.1: *The complete objective function of InfoBERT, which can be decomposed into (a) standard task objective, (b) Information Bottleneck Regularizer, and (c) Local Anchored Feature Regularizer.*

5 Timeline

Here is the proposed timeline of our project.

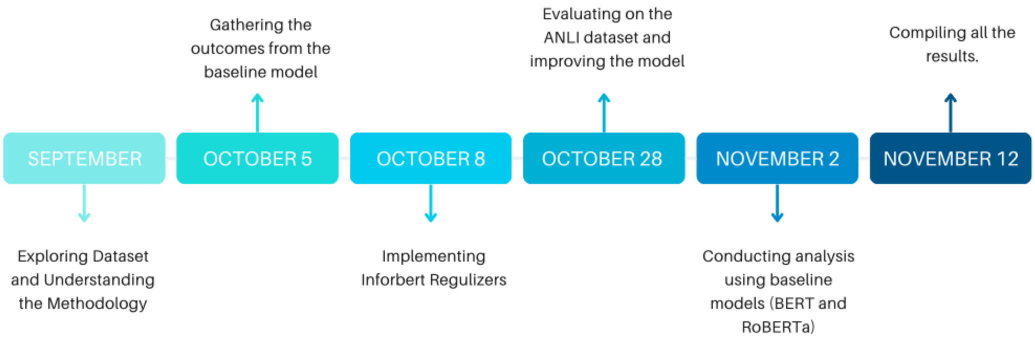
SEPTEMBER
Exploring Dataset and Understanding the Methodology
OCTOBER 5
Gathering the outcomes from the baseline model
OCTOBER 8
Implementing Inforbert Regulizers
OCTOBER 28
Evaluating on the ANLI dataset and improving the model
NOVEMBER 2
Conducting analysis using baseline models (BERT and RoBERTa)
NOVEMBER 12
Compiling all the results.

Figure 5.1: *Timeline for the project*

Bibliography

Alzantot, Moustafa et al. (Oct. 2018). "Generating Natural Language Adversarial Examples". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff et al. Brussels, Belgium: Association for Computational Linguistics, pp. 2890–2896. DOI: 10.18653/v1/D18-1316. URL: https://aclanthology.org/D18-1316.

Cohen, Jeremy M, Elan Rosenfeld, and J. Zico Kolter (2019). *Certified Adversarial Robustness via Randomized Smoothing*. arXiv: 1902.02918 [cs.LG]. URL: https://arxiv.org/abs/1902.02918.

Dvijotham, Krishnamurthy et al. (2018). *Training verified learners with learned verifiers*. arXiv: 1805.10265 [cs.LG]. URL: https://arxiv.org/abs/1805.10265.

Ebrahimi, Javid et al. (July 2018). "HotFlip: White-Box Adversarial Examples for Text Classification". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, pp. 31–36. DOI: 10.18653/v1/P18-2006. URL: https://aclanthology.org/P18-2006.

Gan, Zhe et al. (2020). *Large-Scale Adversarial Training for Vision-and-Language Representation Learning*. arXiv: 2006.06195 [cs.CV]. URL: https://arxiv.org/abs/2006.06195.

Huang, Po-Sen et al. (Nov. 2019). "Achieving Verified Robustness to Symbol Substitutions via Interval Bound Propagation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 4083–4093. DOI: 10.18653/v1/D19-1419. URL: https://aclanthology.org/D19-1419.

Jia, Robin and Percy Liang (Sept. 2017). "Adversarial Examples for Evaluating Reading Comprehension Systems". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2021–2031. DOI: 10.18653/v1/D17-1215. URL: https://aclanthology.org/D17-1215.

Jiang, Haoming et al. (July 2020). "SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 2177–2190. DOI: 10.18653/v1/2020.acl-main.197. URL: https://aclanthology.org/2020.acl-main.197.

Jin, Di et al. (2020). *Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment*. arXiv: 1907.11932 [cs.CL]. URL: https://arxiv.org/abs/1907.11932.

Liu, Xiaodong et al. (2020). *Adversarial Training for Large Neural Language Models*. arXiv: 2004.08994 [cs.CL]. URL: https://arxiv.org/abs/2004.08994.

Ren, Shuhuai et al. (July 2019). "Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 1085–1097. DOI: 10.18653/v1/P19-1103. URL: https://aclanthology.org/P19-1103.

Ye, Mao, Chengyue Gong, and Qiang Liu (July 2020). "SAFER: A Structure-free Approach for Certified Robustness to Adversarial Word Substitutions". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 3465–3475. DOI: 10.18653/v1/2020.acl-main.317. URL: https://aclanthology.org/2020.acl-main.317.

Zang, Yuan et al. (July 2020). "Word-level Textual Adversarial Attacking as Combinatorial Optimization". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 6066–6080. DOI: 10.18653/v1/2020.acl-main.540. URL: https://aclanthology.org/2020.acl-main.540.

Zhu, Chen et al. (2020). *FreeLB: Enhanced Adversarial Training for Natural Language Understanding*. arXiv: 1909.11764 [cs.CL]. URL: https://arxiv.org/abs/1909.11764.

Zhu, Sicheng, Xiao Zhang, and David Evans (2020). *Learning Adversarially Robust Representations via Worst-Case Mutual Information Maximization*. arXiv: 2002.11798 [cs.LG]. URL: https://arxiv.org/abs/2002.11798.