Advanced NLP Project: Adverserial NLI

Prayush Rathore and Yash Agrawal
B. Tech. CLD (3rd Year)
International Institute of Information Technology,
Hyderabad, Telangana

***Abstract*—Natural language inference (NLI) is the task of determining whether a "hypothesis" is true (entailment), false (contradiction), or undetermined (neutral) given a "premise".**

Example: Premise: A man inspects the uniform of a figure in some East Asian countr Label: Contradiction Hypothesis: The man is sleeping.

Approaches used for NLI include earlier symbolic and statistical approaches to more recent deep learning approaches. Benchmark datasets used for NLI include SNLI, MultiNLI, SciTail, among others. You can get hands-on practice on the SNLI task by following this d2l.ai chapter.

We want to implement InfoBERT training, testing and evaluation on the ANLI dataset. Compare the performance against simple BERT-based baselines to show the effectiveness of the robust finetuning and the difficulty of the ANLI dataset

***Index Terms*—Website, workflow management,**

I. Literature Review

A. Adversarial NLI: A New Benchmark for Natural Language Understanding

The paper provides a brand-new, extensive NLI benchmark dataset that was gathered by an iterative, adversarial human-and-model-in-the-loop procedure. They demonstrate that using the additional dataset to train models yields state-of-the-art performance on several common NLI benchmarks while presenting a more challenging task when using the new test set. The research reveals the flaws in the most recent state-of-the-art models and demonstrates that amateur annotators are capable of identifying these flaws. Instead of using a static benchmark that would rapidly saturate, the data gathering approach may be used in a never-ending learning situation, making it a moving target for NLU.

The paper proposes an iterative, adversarial human-and-model-in-the-loop solution for NLU dataset collection that addresses both benchmark longevity and robustness issues. In the first stage, human annotators devise examples that the current best models cannot determine the correct label for. These resulting hard examples—which should expose additional model weaknesses—can be added to the training set and used to train a stronger model. Then subject the strengthened model to the same procedure and collect weaknesses over several rounds. After each round, a new model is trained and set aside a new test set. The process can be iteratively repeated in a never-ending learning (Mitchell et al., 2018) setting, with the model getting stronger and the test set getting harder in each new round. Thus, not only is the resultant dataset harder than existing benchmarks, but this process also yields a "moving post" dynamic target for NLU systems, rather than a static benchmark that will eventually saturate.

B. InfoBERT: Improving Robustness of Language Models from An Information Theoretic Perspective

Large-scale language models, like BERT, have attained cutting-edge performance on a variety of NLP tasks. However, recent research indicates that these BERT-based models are susceptible to textual adversarial assaults. The InfoBERT, a novel learning framework for secure fine-tuning of pre-trained language models, in an effort to approach this issue from an information-theoretic viewpoint. For the purpose of training models, InfoBERT includes two mutual-information-based regularizers: Information Bottleneck regularizer, which reduces noisy mutual information between the input and the feature representation; and (ii) a Robust Feature regularizer, which boosts mutual information between local robust features and global features. In both conventional and adversarial training, we offer a systematic approach for conceptually analysing and enhancing the robustness of representation learning for language models.

Extensive trials show that InfoBERT delivers state-of-the-art resilient accuracy on Natural Language Inference (NLI) and Question Answering (QA) tasks over a variety of hostile datasets.

Self-supervised representation learning pre-trains good feature extractors from massive unlabeled data, which show promising transferability to various downstream tasks. Recent success includes large-scale pre-trained language models (e.g., BERT, RoBERTa, and GPT-3, which have advanced state of the art over a wide range of NLP tasks such as NLI and QA, even surpassing human performance. Specifically, in the computer vision domain, many studies have shown that self-supervised representation learning is essentially solving the problem of maximizing the mutual information (MI) I(X; T) between the input X and the representation T .

II. Dataset Used

The primary aim of the first paper is to create a new large-scale NLI benchmark on which current state-of-the-art models fail. This constitutes a new target for the field to work towards, and can elucidate model capabilities and limitations. As noted, however, static benchmarks do not last very long these days. If continuously deployed, the data collection procedure they introduce here can pose a dynamic challenge that allows for never-ending learning.

gold_label	sentence1	sentence2
-	785	785
contradiction	183187	183185
entailment	183416	183414
neutral	182764	182762

Fig. 1.

The following adversarial datasets and adversarial attacks will be used by us to evaluate the robustness of InfoBERT and baselines. (I) Adversarial NLI (ANLI) (Nie et al., 2020) is a large-scale NLI benchmark, collected via an iterative, adversarial, human-and-model-in-the-loop procedure to attack BERT and RoBERTa. ANLI dataset is a strong adversarial dataset which can easily reduce the accuracy of BERTLarge to 0. (II) Adversarial SQuAD (Jia Liang, 2017) dataset is an adversarial QA benchmark dataset generated by a set of hand-crafted rules and refined by crowdsourcing. Since adversarial training data is not provided, we fine-tune RoBERTaLarge on benign SQuAD training data (Rajpurkar et al., 2016) only, and test the models on both benign and adversarial test sets. (III TextFooler (Jin et al., 2020) is the state-of-the-art word-level adversarial attack method to generate adversarial examples. To create an adversarial evaluation dataset, we sam- pled 1, 000 examples from the test sets of SNLI and MNLI respectively, and run TextFooler against BERTLarge and RoBERTaLarge to obtain the adversarial text examples.

III. Present Work

We implemented a BERT based model on the available SNLI dataset by Stanford, The model has 109,484,547 trainable parameters. We ran the model for one epoch only for the interin submission. Batch size used is sixteen. We have used the Adam Optimizer and for the loss function we are using cross-entropy loss function.
Train Loss: 0.390 — Train Acc: 85.54
Test Loss: 0.294 — Test Acc: 89.55

IV. Conclusion

We aim at achieving a better performance compared to the current SoTA models. The results of the previous state of the art are given in the picture above.

Proposed Timeline End of October: Collection of Data, Dataset Preprocessing, Developing and Training of a Model. Mid-November: Writing a report and finetuning the model.

Model	SNLI-Hard	NLI Stress Tests					
		AT (m/mm)	NR	LN (m/mm)	NG (m/mm)	WO (m/mm)	SE (m/mm)
Previous models	72.7	14.4 / 10.2	28.8	58.7 / 59.4	48.8 / 46.6	50.0 / 50.2	58.3 / 59.4
BERT (All)	82.3	75.0 / 72.9	65.8	84.2 / 84.6	64.9 / 64.4	61.6 / 60.6	78.3 / 78.3
XLNet (All)	83.5	88.2 / 87.1	85.4	87.5 / 87.5	59.9 / 60.0	68.7 / 66.1	84.3 / 84.4
RoBERTa (S+M+F)	84.5	81.6 / 77.2	62.1	88.0 / 88.5	61.9 / 61.9	67.9 / 66.2	86.2 / 86.5
RoBERTa (All)	84.7	85.9 / 82.1	80.6	88.4 / 88.5	62.2 / 61.9	67.4 / 65.6	86.3 / 86.7

Fig. 2.