Experiment Protocol for CS224U, Spring 2020.

Abhishek Goswami
Microsoft / Redmond, WA
agoswami@microsoft.com

Abstract

This document sets the experiment protocol for CS224U, Spring 2020 quarter.

1 Hypothesis

Natural language inference (NLI) is the task of determining if a natural language hypothesis can be inferred from a given premise in a justifiable manner. Existing models perform well at standard evaluations for NLI, achieving impressive results in leaderboards such as GLUE. However, a growing body of evidence (McCoy et al., 2019; Glockner et al., 2018) shows that state-of-the-art models learn to exploit spurious statistical patterns in datasets instead of learning meaning in the flexible and generalizable way. NLI adversarial evaluations, where existing state-of-the art NLI systems are evaluated on a completely unseen new test dataset, further exposes these concerns. State-of-the art NLI systems perform quite poorly in the adversarial evaluation setting (Nie et al., 2019), reflecting that such systems do not represent true competence at natural language understanding. This begs the question: If a system fails an NLI adversarial evaluation, is it a failing of the model or of the dataset used to develop the model?

In this work, we hypothesize that the failings in the adversarial evaluation setting come from the dataset used to develop the model. Standard evaluations in NLI typically use datasets generated using a single generation process. We argue that these standalone, per-dataset generation processes encode latent signals in the premise/hypothesis construction. This leads to favorable results in the standard evaluation setting, but fail spectacularly in the the adversarial evaluation setting. We believe this can be improved by using a mix of multiple weakly-supervised datasets during the training process. This setting enables us to leverage several related sentence pair datasets, while avoiding per-dataset statistical concerns.

2 Data

In this section we provide a description of the datasets that we use in this project. We also describe also how we cast some of these datasets for the NLI task, as a form of weak supervision. Table 1 gives a summary of the distribution of labels across different datasets.

2.1 Test Set

For our test set we use the Adverserial NLI dataset, as described in (Nie et al., 2019). More specifically we choose the *test set from Round1* as our test set. This test set is referred to as ANLI-A1 in this project. By construction, this test set poses a serious challenge to our models. The examples in this dataset were verified as correct by human annotators. But a state-of-

Categories	Training Datasets							Dev Set	Test Set
	MNLI	MRPC	QQP	STSB	QNLI	RTE	WNLI	MNLI-(m/mm)	ANLI-A1
Contradiction	131k	1194	229k	1773	52k	1241	312	3,213 / 3,240	333
Entailment	131k	2474	134k	1052	52k	1249	323	3,479 / 3,463	334
Neutral	131k	-	-	2924	-	-	-	3,123 / 3,129	333

Table 1: Distribution of labels across different datasets

the-art $\text{BERT}_{\text{LARGE}}$ model trained on SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2017) datasets got them all wrong. Also, in the spirit of adversarial evaluation, we do not consider any other ANLI datasets during training.

2.2 Dev Set

For our dev set we use both the matched and mismatched dev sets which are part of the MNLI dataset. We chose this as our dev set because it contains all the three different labels of interest. Please do note that since we are interested in the adversarial evaluation setting, the choice of dev set is purely arbitrary.

2.3 Train Set

We give a brief summary of the different datasets that we consider in our mix for training. The common theme for all these datasets is that they lend themselves to pairwise text classification. We argue that each of them can be cast for the NLI setting.

MNLI: The Multi-Genre Natural Language Inference Corpus (Williams et al., 2018), is a crowdsourced collection of sentence pairs with textual entailment annotations. Given a premise sentence and a hypothesis sentence, the task is to predict whether the premise entails the hypothesis (`entailment`), contradicts the hypothesis (`contradiction`), or neither (`neutral`). Given the presence of all the 3 categories of interest, it becomes a natural choice to be used for training.

MRPC: The Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005) is a corpus of sentence pairs automatically extracted from online news sources, with human annotations for whether the sentences in the pair are semantically equivalent. We feel this dataset can be used for NLI as a form of weak supervision, with semantically equivalent sentences considered as `entailment`, and `contradiction` otherwise.

QQP: The Quora Question Pairs2 (Iyer et al., 2017) dataset is a collection of question pairs from the community question-answering website Quora. The task is to determine whether a pair of questions are semantically equivalent. We feel this dataset can be used for NLI as a form of weak supervision, with semantically equivalent questions considered as `entailment`, and `contradiction` otherwise.

STSB: The Semantic Textual Similarity Benchmark (Cer et al., 2017) is a collection of sentence pairs drawn from news headlines, video and image captions, and natural language inference data. Each pair is human-annotated with a similarity score from 1 to 5. In order to use this for NLI, we do the following pre-processing : Pairs of sentences with similarity score in the range [1,2) , [2, 4] and (4, 5] are considered as `contradiction`, `neutral` and `entailment` respectively.

QNLI: This is modified version of the

Stanford Question Answering Dataset (Rajpurkar et al., 2018). The modification aims to determine whether the context sentence contains the answer to the question. For a given (context, question) pair, if it is a relevant questions (i.e. can be answered) then the pair is treated as `entailment`, and `contradiction` otherwise.

RTE: The Recognizing Textual Entailment (RTE) datasets come from a series of annual textual entailment challenges. The dataset we use combines the data from RTE1 (Dagan et al., 2005), RTE2 (Bar-Haim et al., 2006), RTE3 (Giampiccolo et al., 2007), and RTE5 (Bentivogli et al., 2009).

WNLI: This is modified version of the Winograd Schema Challenge (Levesque et al., 2012), the modification being that sentence pairs are constructed by replacing the ambiguous pronoun with each possible referent.

3 Metrics

We measure our model performance using the accuracy metric. As shown in Table 1, our dev and test sets have a uniform distribution across all the three labels. We believe this justifies using accuracy as the evaluation metric.

4 Models

In this section we give a brief summary of the different models we consider in this study.

4.1 Random

This is a simplest case, where the model randomly selects one of the three classes: `contradiction`, `neutral` `entailment`

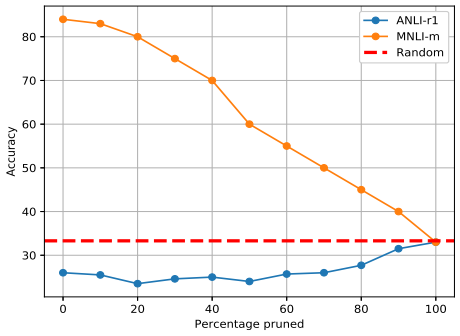

Figure 1: Accuracy as a function of attention heads. $\text{BERT}_{\text{BASE}}$ model trained on MNLI only.

4.2 Baseline

Our baseline model is a hypothesis-only simple RNN classifier. Hypothesis-only baselines for NLI tasks can be remarkably robust, and hence we chose it as our baseline model. For the embedding layer, we use 50 dimensional Glove (Pennington et al., 2014) embeddings. We use a uni-directional LSTM with a hidden dimension of 50.

4.3 BERT

BERT (Devlin et al., 2019) is one of the Transformer-based models that we include in our study. We use `bert-base-uncased` which is a 12-layer, 768-hidden, 12-heads, 110M parameters model.

4.4 RoBERTa

RoBERTa (Liu et al., 2019) is the second Transformer-based models that we include in our study. We use roberta-base, which is a 12-layer, 768-hidden, 12-heads, 125M parameters model.

5 General Reasoning

Our hypothesis is that that the failings in NLI adversarial evaluation come from the use of singular datasets used to develop the model.

System	Training Data Ratio Used							Dev Acc	Test Acc
	MNLI	MRPC	QQP	STSB	QNLI	RTE	WNLI	MNLI-(m/mm)	ANLI-A1
Random	1.0	-	-	-	-	-	-	33.8 / 33.2	33.3
Baseline	1.0	-	-	-	-	-	-	50.3 / 49.6	28.9
$\text{BERT}_{\text{BASE}}$	1.0	-	-	-	-	-	-	84.1 / 84.4	26.0
$\text{RoBERTa}_{\text{BASE}}$	1.0	-	-	-	-	-	-	87.7 / 87.6	28.5
$\text{RoBERTa}_{\text{LARGE}}$	1.0	-	-	-	-	-	-	92.0 / 90.0	46.0
$\text{BERT}_{\text{BASE}}$ (weak supervision)	0.1	0.1	0.1	-	-	-	-	76.9 / 77.2	23.2
	0.1	0.1	0.1	0.1	0.1	0.1	0.1	77.3 / 77.6	25.8
	1.0	1.0	1.0	1.0	1.0	1.0	1.0	84.5 / 83.8	27.1
$\text{RoBERTa}_{\text{BASE}}$ (weak supervision)	0.1	0.1	0.1	-	-	-	-	83.6 / 84.1	26.0
	0.1	0.1	0.1	0.1	0.1	0.1	0.1	83.6 / 84.3	29.3
	1.0	1.0	1.0	1.0	1.0	1.0	1.0	87.5 / 87.5	31.3

Table 2: Evaluation results on the Dev/Test sets

Categories	$\text{BERT}_{\text{BASE}}$	$\text{RoBERTa}_{\text{BASE}}$
Contradiction	369	353
Entailment	303	308
Neutral	328	339

Table 3: Distribution of predicted labels on ANLI-A1. Models trained on MNLI only.

To avoid this we introduce a wide range of datasets as part of our train set. On the model side, we have two state-of-the art models: $\text{BERT}_{\text{BASE}}$ and $\text{RoBERTa}_{\text{BASE}}$ which perform exceedingly well in standard NLI tasks. By diversifying our training data, we aim to show that these two models can also do well in the NLI adversarial evaluation setting.

6 Progress Summary

In this section, we report the progress we have made so far. Table 2 shows a comprehensive view of the experiments done so far.

6.1 Training on MNLI only

A random guessing strategy gives a accuracy of 33.3% on both the Dev and Test sets which is intuitive given that we have 3 labels which are uniformly distributed (See Section 2).

Compared to Random, our Baseline model does better on the Dev set indicating it has picked up signal from the MNLI training data. However, it does worse than Random in the adversarial evaluation setting when evaluated on the ANLI-A1 test set. We see similar behaviour for both the 'simple' Transformer-based models, . When trained on MNLI data both $\text{BERT}_{\text{BASE}}$ and $\text{RoBERTa}_{\text{BASE}}$ do remarkably well on the Dev set, but do quite poorly on the ANLI-A1 test set. In this setting, only $\text{RoBERTa}_{\text{LARGE}}$ is able to outperform Random on the Test set.

6.2 Analyzing Transformer models on Adversarial data

We try to understand what Transformed-based models are doing from two perspectives (a) Error analysis and (b) Role of the attention heads. For this we continue to use models trained on MNLI only.

Error Analysis: We look at some examples where *both* $\text{BERT}_{\text{BASE}}$ and $\text{RoBERTa}_{\text{BASE}}$ are wrong, while being in perfect agreement with each other. Table 4 shows one example from each such category. This leads to a few observations:

- Adversaries are exploiting numerical abilities, including in some cases the inability of the models to do math.

Premise	**Hypothesis**	**gold_label**	**BERT$_{\text{BASE}}$**	**RoBERTa$_{\text{BASE}}$**
Julian Peter McDonald Clary (born 25 May 1959) is an English comedian and novelist. Openly gay, Clary began appearing on television in the mid-1980s and became known for his deliberately stereotypical camp style. Since then he has also acted in films, television and stage productions, and was the winner of Celebrity Big Brother 10ïn 2012.	Born on May 25, 1958 Julian Clary knew he would be gay.	contradiction	entailment	entailment
Bassingham is a village and civil parish in the North Kesteven district of Lincolnshire, England. The population of the civil parish at the 2011 census was 1,425. The village is situated approximately 8 mi south-west from the city and county town of Lincoln.	Bassingham is a village and civil parish in the North Kesteven district of Lincolnshire, England and only has 1,435 citizens as of the 2011 census.	contradiction	neutral	neutral
Idrees Kenyatta Walker (born February 1, 1979) is a former professional American football player who was an offensive tackle in the National Football League (NFL) for six seasons. Walker played college football for the University of Florida. A first-round pick in the 2001 NFL Draft, he played professionally for the Tampa Bay Buccaneers of the NFL.	Kenyatta Walker did not play football at Florida State.	entailment	contradiction	contradiction
The 2005 Big East Men's Basketball Championship was played from March 9 to March 12, 2005. The tournament took place at Madison Square Garden in New York City. The Syracuse Orange won the tournament and were awarded an automatic bid to the 2005 NCAA Men's Division I Basketball Tournament.	The tournament took place over four days.	entailment	neutral	neutral
Puss in Boots is an action game based on the DreamWorks Animation SKG movie of the same name. It was developed by Blitz Games, and released by THQ on October 25, 2011 for Xbox 360, PlayStation 3, Wii and Nintendo DS. It features support for Kinect and PlayStation Move on the respective platforms. It was released on October 25, 2011 in North America and December 2 for Europe.	Dreamworks released a movie in 2012	neutral	contradiction	contradiction
Binani Industries Ltd is an Indian business group based in Mumbai. It is a 143-year old business conglomerate and belongs to the legendary Braj Binani Group. The business portfolio of Binani Industries includes sectors like cement, zinc, glass-fiber, and downstream composite products.	Braj Binani Group has enjoyed ownership of Binani Industries Ltd since its creation.	neutral	entailment	entailment

Table 4: Examples where *both* BERT$_{\text{BASE}}$ and RoBERTa$_{\text{BASE}}$ are wrong, while being in perfect agreement with each other. Models trained on MNLI only.

- Both models seem to biased towards predicting the `contradiction` label, with the number of predicted `contradiction` labels being almost 20% more than the number of `entailment` labels.

Attention heads: We look at the role of attention heads for both the standard evaluation setting (using MNLI-m as the target dataset) and for the adversarial evaluation setting (using ANLI-A1 as the target dataset). In (Michel et al., 2019) the authors note that at test time, many attention heads can be re- moved individually without impacting model performance. In Section 6.1 we noted that a $\text{BERT}_{\text{BASE}}$ trained on MNLI data did did worse than Random on the ANLI-A1 test set. Interestingly, we saw that we had to prune almost 90% of the attention heads of the model for it to match Random. Figure 1 accuracy of a $\text{BERT}_{\text{BASE}}$ model as a function of the number of attention heads. The $\text{BERT}_{\text{BASE}}$ model was trained on MNLI only data.

6.3 Training on multiple data sources

In line with our stated hypothesis, we now mix multiple datasets to construct our train data. The hope is that this will enable us to capture a diverse range of latent signals from multiple data generation processes. For efficiency purposes, we currently sample from the data ratio to gain an intuition on how well our models are doing. In this setting, we only report the results for the $\text{BERT}_{\text{BASE}}$ and $\text{RoBERTa}_{\text{BASE}}$ models

Unfortunately, so far our results are in line with the observations in Section 6.1. Both the transformer models are doing remarkably well on the Dev set, but poorly on the ANLI-A1 Test set.

6.4 Advice

Any suggestions on how to further improve the models would be highly welcome. Specifically in adressing the limitation of models with respect to their numerical abilities as noted above in Section 6.2

References

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the second PASCAL challenges workshop on recognising textual entailment*, volume 6, pages 6–4. Venice.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. *arXiv preprint arXiv:1805.02266*.

Shankar Iyer, Nikhil Dandekar, and Kornel Csernai. 2017. First quora dataset release: Question pairs.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

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

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, pages 14014–14024.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.