Assignment 2: Adversarial Augmentation of FEVER Dataset

Matteo Manias 1822363

Abstract

This report for the second assignments focuses on achieving natural language inference robustness on the NLI task on and adversarial augmentation of the FEVER dataset, using a transformer-based model. The report discusses the methodology considered and implemented to augment the standard FEVER dataset with adversarial sample, the experiments conducted, and the results obtained.
The techniques used to this end include word substitution, and premise and hypothesis pair generation.
The results obtained show that a generic model is able to learn more robust features and generalizes better to unseen data by training on the augmented dataset.

1 FEVER Dataset

The dataset used for trainig and test is primarly a NLI dataset, the generic sample format is as follows:

- **Premise**: An input sentence that serves as the context
- **Hypothesis**: An sentence that may or may not be entailed to the premise
- **Label**: The label of the pair, which can be either entailment, neutral or contradiction
- **wsd**: The word substitution indexes for common wsd libraries (e.g. WordNet, Nltk)
- **srl**: The semantic role labeling indexes for common srl libraries (e.g. BabelNet)

1.1 Data split and distribution

The dataset is split into:

- **Train**: 51.1k samples
- **Validation**: 2.29k samples
- **Test**: 2.29k samples

2 Model Architecture

The model used for evaluation of the dataset augmentation is a transformer-based model, specifically the BERT model, available in the HuggingFace library 'bert base uncased', which is a 110M paramenter model.
The pretrined model is fine-tuned for the classification with layer freezing

3 Data Augmentation

The data augmentation implemented:

- **Word Substitution**: Substitution of words in the sentence with synonyms, hyponyms, hypernyms.
- **Premise and Hypothesis Pair Generation**: A new dataset sample is generated by combining the premise and hypothesis of two different entries.

Other techniques that have been considered and motivation for not implementing them:

- **Back Translation**: Translation of the sentence to another language and back to the original language, widely used technique but not introduced in this study case as it uses external resources.
- **Random Noise Injection**: Addition of random noise to the sentence. Standard augmentation technique, very simple to implement with existing libraries.
- **Srl schemas manipulation(e.g. Actor patient inversion)**: Inversion of the actor and patient in the sentence. Powerful technique but very difficult to implement algorithmically for a widre range of schemas to be effective as it affects a limited number of samples.

3.1 Word Substitution using the wsd indexes

While the problem of substituting words in a sentence with synonyms, hypernyms ect. may seem trivial given that one can leverage existing libraries and indexes, like Wordnet,Nltk and BabelNet that have built in functions for this purpose, a careful examination of this approach reveals a problem that undermines the effectiveness of, with regard to the context of the sentence, blind substitution: often the meaning of the augmented sentence is not coherent with the original.

To address this issue, the approach presented in this study compares the augmented sentence with candidates max number of words for substitution per type (synonym,hypernym,hyponym), to the original sentence and based on a tunable similarity score threshold, produces a coherent augmented sentence.

Eventough this approach is more computationally expensive and cannot guarantee that the augmented sentence is always coherent, it is more effective than blind substitution at generating a qualitatively coherent sentence. The resulting dataset is also enhanced with aumentation info of the words that have been substituted, to allow for further analysis and evaluation of the augmentation.

The dataset is augmented from the original training set of 51.1k samples, and augmented to a number of 95.8k samples. Note: this may be excessive and lead to overfitting. The result on test set is shown in Figure 3.

3.2 Premise and Hypothesis Pair Generation

Generating new premise and hyphosis pairs is a powerful technique to augmente the dataset, as it provides the model contrastive examples, e.g the same premise with the negation of the hypothesis, and neutral examples, e.g. the same premise with a different hypothesis.

this allows the model to learn a more robut representation of the entailment logic and generalize better to unseen data.

This technique may be also useful to balance a dataset that is skewed towards a specific label, in this case the dataset is heavily skewed towards the entailment label (60.9 % of the samples)

Generation of **NEUTRAL** samples is simply achieved by randomly selecting two samples from the dataset and combining their premise and hypothesis and assigning the label NEUTRAL.

Generation of **CONTRADICTION** samples is achieved by selecting samples with the **ENTAILMENT** label, negating the hypothesis and assiging the label accordingly. Negating a sentence is however a non-trivial task, following the considerations expressed in the previous section, The new negated sentence is obtained by leveraging similar methods of context-aware substitution using the `negated` library.

The dataset is augmented from the original training set of 51.1k samples, and augmented to a number of 78.8k samples. Note: the number of negative and neutral new pair is chosen to balance the dataset. The result on test set is shown in Figure 3.

4 Experiments

The effectiveness of the augmentation strategies are tested by fine-tuning the BERT model on the augmented dataset and evaluating the model on the adversarial test set. The training and testing pipeline is implemented using the HuggingFace API that provides a simple and efficient way to fine-tune and evaluate transformer-based models. The evaluation metrics used are the standard accuracy, precision, recall and F1 score.

4.1 Fine-tuning on the word substitution augmented dataset results

Results in the table 1

Metric	Standard Dataset	Augmented Dataset
Test Loss	1.2669	1.2446
Test Accuracy	0.3463	0.3383
Test Precision	0.1199	0.3845
Test F1	0.1782	0.1863
Test Recall	0.3463	0.3383

Table 1: Comparison of model performance on the adversarial test set between finetuning on the standard dataset and the augmented dataset. Only some metrics are improved by the augmentation, namely the precision and F1 score.

4.2 Fine-tuning on the premise and hypothesis pair generation augmented dataset results

Result in the table 1

5 Conclusion

The results show that the two techniques of data augmentation are effective at improving the model's performance on adversarial samples.

The pair generation technique is more effective

Metric	Standard Dataset	Augmented Dataset
Test Loss	1.2669	1.0705
Test Accuracy	0.3463	0.4065
Test Precision	0.1199	0.4436
Test F1	0.1782	0.3814
Test Recall	0.3463	0.4065

Table 2: Comparison of model performance on the adversarial test set between finetuning on the standard dataset and the augmented dataset. Accuracy and precision are improved significantly by the augmentation. F1 and Recall scores are also impoved.

than the word substitution technique, as it provides the model with more diverse and contrastive examples, the approch here presented serves only as a study of different individual augmentation techniques, a more effective approach must combine multiple techniques to achieve the best results.

Augmentation examples	
alteration	**synonym,hypernym,hyponym substitution**
hypothesis	Roman Atwood is a content creator.
new hypothesis	Roman Atwood is a content designer.
augmentation info	(‘original_word‘: ‘creator‘, ‘substituted_word‘: ‘designer‘, ‘similarity_score‘: 0.9385678172111511)
alteration	**synonym,hypernym,hyponym substitution**
hypothesis	The Boston Celtics play their home games at TD Garden.
new hypothesis	The Boston Celtics compete their home games at TD Garden.
augmentation info	(’original_word’: ’play’, ’substituted_word’: ’compete’, ’similarity_score’: 0.974838376045227)
alteration	**synonym,hypernym,hyponym substitution**
hypothesis	Ryan Seacrest is a person.
new hypothesis	Ryan Seacrest is an individual.
augmentation info	(‘original_word‘: ‘person‘, ‘substituted_word‘: ‘individual‘, ‘similarity_score‘: 0.9043632745742798)

Figure 1: Examples of word substitution augmentations, the substituted word is chosen among a set of candidates retrieved from the wsd indexes and chosen based on a similarity score threshold computed on the whole sentence.

Augmentation examples	
alteration	**Neutral Pair generation**
hypothesis	Stranger than Fiction is a 2006 American fantasy comedy-drama film directed by Marc Forster , produced by Lindsay Doran , and written by Zach Helm .
new hypothesis	Selena recorded music.
alteration	**Neutral Pair generation**
hypothesis	Ryan John Seacrest (born December 24 , 1974) is an American radio personality , television host and producer . Seacrest began co-hosting Live with Kelly and Ryan on a permanent basis May 1 , 2017 . He received Emmy Award nominations for American Idol , and won an Emmy for producing Jamie Oliver ’s Food Revolution .
new hypothesis	Roman Atwood is a content creator.

Figure 2: Example of neutral pair generation from the existing sentences in the dataset, no further modification is needed at a word level.

Augmentation examples	
alteration	**Negative Pair generation**
hypothesis	Roman Atwood . He is best known for his vlogs , where he posts updates about his life on a daily basis . His vlogging channel , “ RomanAtwoodVlogs ” , has a total of 3.3 billion views and 11.9 million subscribers . He also has another YouTube channel called “ RomanAtwood ” , where he posts pranks .
new hypothesis	Roman Atwood isn’t a content creator.
alteration	**Neegative Pair generation**
hypothesis	Chester Bennington . He is best known as the lead vocalist of rock bands Linkin Park , Dead by Sunrise , and live rock cover band Bucket of Weenies . Bennington was the lead vocalist for Stone Temple Pilots from 2013 to 2015 .
new hypothesis	Chester Bennington is not a singer.

Figure 3: Examples of negative pair generation, the hypothesis is negated by negating the verb in the sentence and the label is changed to contradiction.