THE UNIVERSITY OF TEXAS AT AUSTIN

Mitigating Dataset Artifacts Through Automatic Contextual Data Augmentation and Learning Optimization

Michail Mersinias
Panagiotis Valvis

1 Problem

Natural language inference (NLI) is to determine if a natural language hypothesis can justifiably be inferred from a natural language premise. It may be inferred as Entailment (ENT), Neutral (NEU) or Contradiction (CON).

Recent research shows that a model may achieve high performance on a dataset by learning spurious correlations, which are called dataset artifacts, but it is then expected to fail in settings where these artifacts are not present, such as adversarial cases.

2 Solution

We propose a combined approach which comprises three techniques towards mitigating dataset artifacts:

- A novel data augmentation procedure, named ACDA: Automatic Contextual Data Augmentation.
- Contrastive Learning.
- Hybrid loss function.

3 ACDA

ACDA: Automatic Contextual Data Augmentation.

A data augmentation procedure which scans the hypothesis sentence for nouns, and queries WordNet synsets for a replacement word. It then swaps each one of the nouns at a time and composes new examples using our label generation rules.

ACDA yields a high number of new training examples from the most problematic areas of the decision boundary, which can now be used as part of training to incentivise the model against the reliance on dataset artifacts.

The resulting augmentation benefits from being both fully automatic, as it does not require manual writing of new hypothesis or label annotation, while at the same time being non-trivial. It is also computationally efficient, as it adds no overhead to the training cost.

4 Contrastive Learning

A learning optimization technique which further incentivise the model to learn the nuances of the decision boundary. For this purpose, the model has to see instance bundles during training, that is, examples that are close together and belong to a specific area of the decision boundary in the same training batch.

Since ACDA places the augmented examples right after each original one, the dataset batches provided to the model in each iteration will consist of some number of original examples and their augmentations. This way, we manage to have a dataset consisting of multiple instance bundles and therefore, we gain the maximum benefit from contrastive learning.

5 Hybrid Loss

The contrastive learning optimization technique re-focuses training in the localities of the current batch, but there lies the danger of the model learning to overfit these localities, while not being able to correctly classify examples that it has not seen and are further apart in decision space.

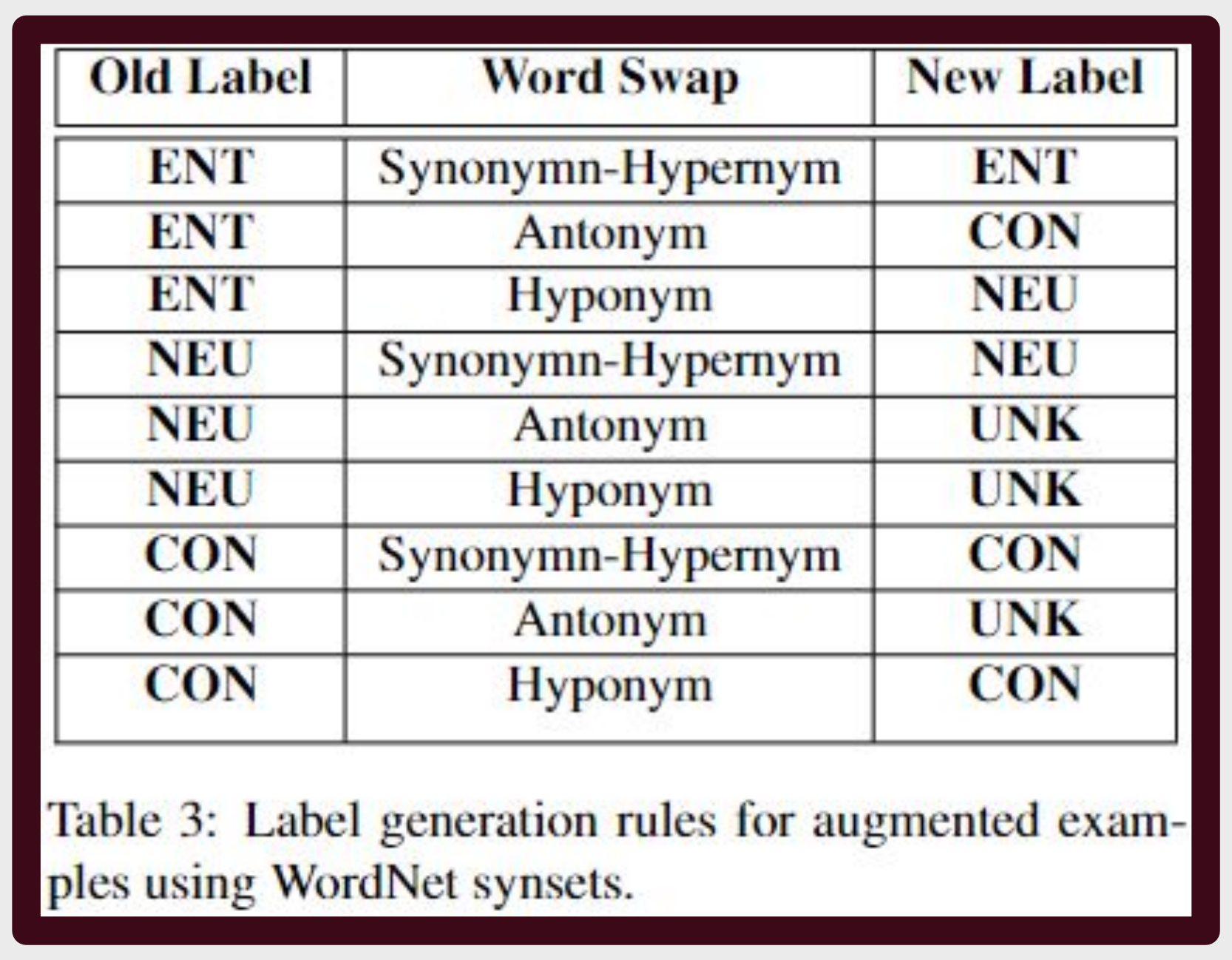
Our Hybrid Loss combines Cross Entropy Loss and NNL Loss in a weighted average manner. This way, we manage to retain the advantages of both loss functions. The Cross Entropy Loss ensures that part of the loss signal will be directly relevant to the shortcomings of the model in the localities of the decision boundary, enabling contrastive learning, while the NLL Loss will incentivise generalization in areas that the model has not seen, learning rules that can only be inferred by looking at unrelated examples.

6 Synergy

ACDA generates a dataset which is up to 10 times larger than the original one, while remaining computationally efficient. It generates adversarial examples which form instance bundles in areas of the decision boundary.

Contrastive Learning takes advantage of the instance bundles generated by ACDA in order to optimize the learning process.

Hybrid Loss further optimizes the learning process by ensuring we retain the benefits of Contrastive Learning without losing learning generalization.



Old Label	Word Swap	New Label
ENT	Synonymn-Hypernym	ENT
ENT	Antonym	CON
ENT	Hyponym	NEU
NEU	Synonymn-Hypernym	NEU
NEU	Antonym	UNK
NEU	Hyponym	UNK
CON	Synonymn-Hypernym	CON
CON	Antonym	UNK
CON	Hyponym	CON

Table 3: Label generation rules for augmented examples using WordNet synsets.

7 Results

(SNLI)	BASELINE	ACDA-BOOSTED
BERT-TINY	78.86	82.01
BERT-MINI	85.06	86.79
BERT-SMALL	87.27	87.90
BERT-MEDIUM	88.92	89.01
ELECTRA-SMALL	89.02	89.82

(MNLI)	BASELINE	ACDA-BOOSTED
BERT-TINY	65.24	69.06
BERT-MINI	72.54	75.22
BERT-SMALL	77.02	78.57
BERT-MEDIUM	80.20	80.39
ELECTRA-SMALL	81.16	81.53

The acda-boosted models consistently outperform the respective fine-tuned baseline models across all datasets. Particularly the more lightweight models exhibit the largest gains of up to 3.18% and 3.82% for the SNLI and the MNLI datasets respectively.

For the adversarial dataset, the best performing acda-boosted model (ELECTRA-Small) resulted in a better and more robust behavior, with even larger gains in performance.

8 Technologies

Python 3.9: The programming language that we used.

Wordnet: The large lexical database of English that we used to create our label generation rules for ACDA.

HuggingFace Transformers: The framework that we used to download, fine-tune and train the pre-trained language models.

Pre-trained Models: The pre-trained models that we used were BERT-Tiny, Bert-Mini, BERT-Small, BERT-Medium and ELECTRA-Small.

Stanford Natural Language Inference (SNLI) dataset: A collection of 570000 human-written English sentence pairs manually labeled for balanced classification.

Multi-Genre Natural Language Inference (MNLI) dataset: A crowd-sourced collection of 433000 sentence pairs which cover a range of genres of spoken and written text, and support a distinctive cross-genre generalization evaluation

Adversarial dataset: A small set of hand-annotated adversarial examples which we created.