DS-GA 1012: Breaking Numerical Reasoning in NLI

Jungkyu (JP) Park
New York University
jp4989@nyu.edu

Grace Han
New York University
jh5990@nyu.edu

Yanchao Ni
New York University
yn811@nyu.edu

Mingsi Long
New York University
ml5893@nyu.edu

Abstract

We show that SoTA NLI models fail to perform well on our adversarial test set that involve numerical reasoning, and attempt to achieve better generalization on this data by training on augmented data and modified number word embeddings. We show that data augmentation improves performance on data without addition but not data with addition. Modification of embedding has negligible effect.

1 Introduction

Our preliminary testing shows that state-of-the-art natural language inference system fails to perform numerical reasoning on an adversarial test set we generated. For example, an ESIM model which predicts entailment for pairs which look like ('A man holds two children in his arms.', 'Three humans together.') still predicts entailment for modified pairs which look like ('A man holds two children in his arms.', 'Four humans together.'). We observe below 10% accuracy on such modified test set of neutral examples. We hypothesize that this failure to generalize is due to (1) the lack of training data that involve numerical reasoning (addition in particular) and (2) the lack of structure in word embedding for rare number words (i.e. averaging two number-word embeddings resulting in an embedding for the average of the two numbers) which might prevent models from performing calculations and force them to memorize the observed pairs. Thus, we (1) augment the training data with more sentence pairs that involve numerical reasoning by modifying the existing pairs and (2) modify word embedding in order to better capture the correct relationship between number words. We report our baseline performance on our adversarial test set from models trained on the original SNLI data. Our experiment results are as follows: (i) We observe that performances of SoTA NLI models still break on systematically-generated test set of numerical reasoning. (ii) We show that while GloVe has unclear relationships between number words and word analogies fail, our modified embedding works well with number analogy due to linear relationship between number words. (iii) Data augmentation with pairs of numerical reasoning improves performances on the adversarial test sets while new word embedding has negligible effect. (iv) While performance on adversarial test set without addition improves to 90%, improvement on test set with addition only is relatively small. We speculate that the reason why the SoTA NLI models cannot learn addition after data augmentation and modifying embedding is because there is fundamental limitation of current NLI architectures that cannot learn more complicated tasks than simple pattern matching.

The code is available at https://github.com/jpatrickpark/breaking_numerical_reasoning_nli.

2 Background

SNLI (Bowman et al., 2015) is a dataset for natural language inference task. For each pair of English sentences, the task is to classify the second sentence (hypothesis) as entailment (definitely true), contradiction (definitely false), or neutral (maybe true) given the first sentence (premise) in the context of writing captions for a photo.

The version of GloVe (Pennington et al., 2014) we used is a set of vector representations for words generated as a result of unsupervised learning of language modeling with a large corpus from common crawl with 840B tokens. Semantic relationship between words are preserved in the vector representations such that word analogies can be made by performing subtraction and addition between word embeddings and finding nearest

neighbors (i.e. india to asia is germany to europe).

MacCartney and Manning (2007, 2009) describe how we can apply some logical rules to generate new sentence pairs and the corresponding labels. Minervini and Riedel (2018) uses such generated sentences based on training set for train time data augmentation.

Glockner et al. (2018); Wang et al. (2018) evaluates generalization of neural models with augmented sentences based on the test set. Glockner et al. (2018) generates adversarial test data by taking premises from SNLI and generating multiple hypothesis by replacing a single word into another word that already existed in original SNLI dataset and pretrained word embedding. They show that performances of SoTA models deteriorate significantly (20-30%), suggesting a severe limitation on generalizability. Wang et al. (2018) swaps the premise and hypothesis and claims that for contradiction pairs, the original relationships should still stand. This logic is used for our augmentation for contradiction pairs.

Naik et al. (2018) shows that NLI models fail on a number of reasoning tasks, including numerical reasoning. They create a stress test for numerical reasoning by inserting or changing the phrases "more than" or "less than" in examples that contain quantities. The performances of models drop to less than random. Their analysis shows that models are fooled by syntactic overlap and shallow lexical cues.

We hypothesize that, in order to perform addition, models should be able to store the information about quantity-object pairs with memory component when reading the premise and match the corresponding objects and quantity from the hypothesis. Giulianelli et al. (2018); Trask et al. (2018); Liu et al. (2019) discuss some architectures with memory or arithmetic components that might be better suited for numerical reasoning.

3 Methods

3.1 Generating numerical reasoning data

We now describe how we generated sentence pairs for testing generalization and data augmentation.

We only use SNLI for simplicity and fairness between experiments. We chose SNLI over MNLI since (1) examples in MNLI have more complicated meaning but the numerical reasoning in it is not more difficult than SNLI and (2) SNLI contains 17960 examples which have clear numerical reasoning and are easy to generate data with, while MNLI contains 12239: SNLI contains 1576 entailment pairs with matching plural numerical numbers and 705 entailment pairs with additions, 15029 contradiction pairs with at least one plural numerical word and 650 contradiction pairs with additions while MNLI contains only 2180, 294, 9513, 252 such pairs respectively.

For sentences without addition, we only focus on pairs that have one numerical word in both premise and hypothesis so that the two numbers can be directly compared. For sentences with addition, only the pairs with two numerical words in premise and one numerical word in hypothesis are taken into consideration. The relationship between two sentences are determined based on the pair's *gold_label*. Numerical words that represent value 1 are not considered since replacing such words to bigger numerical words may easily cause issues. "a" is used in many cases to quantify a single object but not all objects that comes after "a" participates in numerical reasoning.

Generated sentence pairs where the numerical values in premise and hypothesis are equal are considered entailment. Pairs where premise contains bigger value than hypothesis are also entailment in some cases but we disregard this case to simplify and avoid any exceptional cases. Sentence pairs where the numerical value in hypothesis is bigger than that in premise for the matching object is considered neutral. We use antonymy to establish contradiction by following the pattern in SNLI dataset. For example, the following sentence pair from original SNLI dataset has gold label of contradiction: "the two boys are swimming with boogie boards" (Premise) and "the two girls are swimming with their floats" (Hypothesis).

We generate data with numerical reasoning by enforcing that there's no overlap of number pairs between training, validation, test set so that we can test generalization to unseen number pairs. For example, for entailment pairs without addition, out of all possible 10 pairs, (8,8) and (2,2) will only be used for validation, and (5,5), (4,4) will only be used for test set for measuring generalization ability to unseen number pairs. For addition, we restrict the number pairs that are already found in training set in augmented training set only. As a result, we have 291152, 4204, 5150 pairs without addition and 52920, 340, 430 pairs with addition in train, dev, test set respectively. The distribu-

		SNLI original	adversarial without addition				adversarial with addition			
Model	Embedding	all	all	entail	cont	neutral	all	entail	cont	neutral
ESIM	GloVe	87.46	24.88	91.32	88.44	8.72	54.31	98.80	97.10	0
ESIM	301D	86.85	21.60	84.47	85.23	5.68	55.24	66.46	97.10	30.36
ESIM	interpolation	87.48	23.86	87.21	88.28	7.74	56.41	66.46	95.65	33.50
BERT	N/A	90.44	22.48	89.04	89.41	5.69	41.72	66.47	97.10	0

Table 1: Baseline test accuracy of models trained with SNLI data only.

tion of labels are approximately 10%, 15%, 75% without addition and 23%, 8%, 69% with addition for entailment, contradiction, and neutral respectively. We upsample to match the number of labels in training time.

3.1.1 Entailment

1. Sentences without addition: Filter entailment pairs that have one numerical word in both premise (*num1*) and hypothesis (*num2*) where *num1* = *num2*. Iterate from 2 to 10 and replace *num1* and *num2* with new values while maintaining *num1* = *num2* relationship.

2. Sentences with addition: Filter entailment pairs that have two numerical words in premise (*num1* and *num2*) and one numerical word in hypothesis (*num3*) where *num1* + *num2* = *num3*. Iterate from 2 to 10 and replace *num1*, *num2* and *num3* with new values while maintaining *num1* + *num2* = *num3* relationship.

3.1.2 Neutral

Do the same as entailment method but keep the number of objects in hypothesis greater than the number of objects in premise.

3.1.3 Contradiction

(1) Swap premise and hypothesis of contradiction pairs with numerical reasoning (2) Modify the pairs we generated for entailment, replace the object in the hypothesis to its antonym. For example, change "two boys" to "two girls".

3.2 Embedding

3.2.1 Relationship between GloVe words

We show that the current number word embeddings in GloVe might not be ideal by visualizing with t-SNE (Maaten and Hinton, 2008) and performing word analogies. As shown in figure 1, we find (i) approximately linear relationship between number words that are used frequently but (ii) no clear connections between clusters (small digits (0-31), bigger digits (32-98) and alphabetical number words) and (iii) that rare number words (purple dots) are equidistant to each other. The word analogies fail as well: candidates for "three to two is eight to x" are four, six, five, nine, seven in increasing order of cosine distance.

We generate two types of modified number embeddings: include an extra 301th dimension for containing normalized number value between 0 and 1, set it to 0 for all other words (301D) and replace embeddings for number words between 1 and 100 to be linear interpolations between embeddings of 1 and 100 from GloVe without an extra dimension (interpolation). Number words in 301D share the same values for first 300 dim which is average of the frequent number words in GloVe. Using either of these embeddings, all analogies automatically succeed with cosine distance of 0 since they are linearly spaced as shown in figure 1.

When using our modified word embedding, we freeze the embedding layer. Furthermore, with 301D, we do not use dropout because we want to make sure the additional 301th dimension is not dropped out.

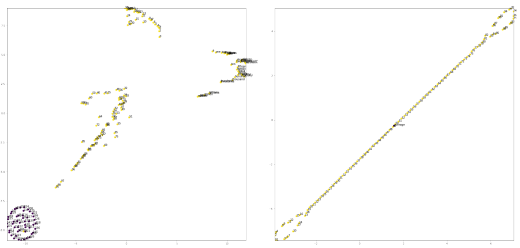

Figure 1: t-SNE plots of number word embeddings. Left: Original GloVe. Right: modified (ours).

3.2.2 Can models learn relationship between numerical words using GloVe?

We generate pairs of numbers with all possible permutations of integers between 0 and 300. The numbers are expressed in words, embedded using GloVe and fed into an LSTM-based model to predict their comparison: greater than, less than, or

	Fine tuned	Mixed
SNLI test	0.44	0.90
Adversarial without addition	1.00	0.99
Adversarial addition	0.66	0.65
Adversarial addition (E)	0.00	0.20
Adversarial addition (N)	1.00	0.87

Table 2: Test accuracy of models trained on augmented data without addition. Entailment (E) and Neutral (N) are separated for testing on augmentation with addition.

equal to. 30% of the number pairs are held out as test set. The model achieves over 99.9% accuracy on test set after convergence on training set, which indicates that although number comparison relationship couldn't be clearly visualized, deep learning models are able to capture it.

3.3 Models

We use E-SIM (Chen et al., 2016) and BERT (Devlin et al., 2018) with the best hyperparameter settings they report.

4 Experiments

Trained on SNLI, test on numerical reasoning with or without addition Our baseline results show that models trained on SNLI only fails on our adversarial test set, with or without addition as shown in table 1. The model ends up predicting entailment for most of the neutral pairs.

Augment without addition, test on addition We first augment the training and validation with our adversarial training set with numerical reasoning without addition and test on addition in order to see if the model becomes capable of addition without specifically augmenting data for it. Results are shown in table 2. Models are able to perform number comparison after training on augmented data. However, on addition examples where numbers in hypothesis are always greater than or equal to numbers in premise, the model only predicts neutral. It seems that models are not doing addition but only comparing numbers.

Augment without addition, test without addition Since fine-tuning seems to only hurt performance to the original test set, we decide to use both original and augmented training sets simultaneously. Results are shown in table 3. While the performance on the original test set is equivalent, performance on adversarial test set improved significantly compared to the baseline as a result of data augmentation. However, no improvement is observed from improved number word embedding.

		SNLI original	adversarial without addition			
Model	Embedding	all	all	entail	cont	neutral
ESIM	GloVe	86.46	95.90	83.10	97.59	96.42
ESIM	301D	86.63	96.93	79.45	95.02	98.42
ESIM	interpolation	86.76	92.90	81.73	96.62	92.94
BERT	N/A	89.73	97.17	89.04	97.11	97.71

Table 3: Test accuracy of models trained with original+augmented data, numerical reasoning without addition.

		SNLI original	adversarial with addition			
Model	Embedding	all	all	entail	cont	neutral
ESIM	GloVe	87.68	69.23	22.75	95.65	100
ESIM	301D	87.57	60.37	0	95.65	100
ESIM	interpolation	87.63	60.60	0	97.10	100
BERT	N/A	90.05	52.91	0	97.10	82.72

Table 4: Test accuracy of models trained with original+augmented data, numerical reasoning with addition only.

Augment with addition, test with addition Results are shown in table 4. There is a slight improvement of the overall performance on the adversarial test set. Without augmentation, the model used to predict entailment for entailment and neutral pairs. As a result of augmentation, the model ended up predicting neutral for all of them, which brings up the overall performance since there are more neutral pairs in the adversarial test set.

5 Discussion

We conclude that data augmentation is enough to achieve satisfactory performance in sentence pairs with all labels without addition and new embedding is not necessarilly better than original GloVe in tackling numerical reasoning. However, neither of them is enough to make models perform well on pairs with addition across all labels.

We speculate that there is a limitation for current natural language inference models because data augmentation helps simple comparison of quantities but not additions. Models capable of simple pattern matching might inherently fall short in being able to count multiple instances of the same kind of objects. This might indicate a need for memory component that enables counting objects and retrieving all relevant information when prompted by hypothesis.

Collaboration statements

JP created starter utility functions for simple data augmentation, generated some sample augmented test set for numerical reasoning involving addition, found out that some SoTA NLI models do not perform well on addition, visualized number word embeddings in GloVe and generated two types of improved number word embeddings, trained and tested ESIM models with SNLI data with and without augmentation of numerical reasoning. Grace explored different structures of numerical reasoning sentences in SNLI, devised methods for data augmentation, and generated augmented dataset for entailment and neutral pairs for sentences where additions exist and doesn't exist. Yanchao set up pipeline for LSTM and BERT training, explored the performance of neural network on number comparison, and trained/tested BERT on augmented data. Mingsi explored sentences involving numerical reasoning in MNLI and SNLI, devised methods for data augmentation, and generated augmented dataset for contradiction pairs for sentences without addition.

References

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, and Hui Jiang. 2016. Enhancing and combining sequential and tree LSTM for natural language inference. *CoRR*, abs/1609.06038.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem H. Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. *CoRR*, abs/1808.08079.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. *CoRR*, abs/1805.02266.

Chunhua Liu, Shan Jiang, Hainan Yu, and Dong Yu. 2019. Multi-turn inference matching network for natural language inference. *CoRR*, abs/1901.02222.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Bill MacCartney and Christopher D Manning. 2007. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200. Association for Computational Linguistics.

Bill MacCartney and Christopher D Manning. 2009. An extended model of natural logic. In *Proceedings of the eighth international conference on computational semantics*, pages 140–156. Association for Computational Linguistics.

Pasquale Minervini and Sebastian Riedel. 2018. Adversarially regularising neural NLI models to integrate logical background knowledge. *CoRR*, abs/1808.08609.

Aakanksha Naik, Abhilasha Ravichander, Norman M. Sadeh, Carolyn Penstein Rosé, and Graham Neubig. 2018. Stress test evaluation for natural language inference. *CoRR*, abs/1806.00692.

Jeffrey Pennington, Richard Socher, and Christopher 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.

Andrew Trask, Felix Hill, Scott Reed, Jack W. Rae, Chris Dyer, and Phil Blunsom. 2018. Neural arithmetic logic units. *CoRR*, abs/1808.00508.

Haohan Wang, Da Sun, and Eric P. Xing. 2018. What if we simply swap the two text fragments? A straightforward yet effective way to test the robustness of methods to confounding signals in nature language inference tasks. *CoRR*, abs/1809.02719.