Combating Adversarial Attacks in NLI Models: Decomposable Attention with Data Augmentation

Jong Hak Choi, Daniel Siu

{eoql11, danielpsiu}@utexas.edu

The University of Texas at Austin Fall 2024

Abstract

Natural Language models have achieved impressive accuracy on standard benchmarks; however, they often rely on superficial heuristics rather than genuine understanding of language. This paper investigates the extent to which the Electra-small model, which is one of the more advanced Natural Language models, depends on such heuristics, and explores potential solutions. First, we modify the model architecture by adapting the decomposable attention to Electra-small to encourage deeper reasoning, and second, we augment the training data with adversarial examples. The custom decomposable attention model improves performance on the lexical overlap heuristic but has limited impact on subsequence and constituent heuristics. In contrast, data augmentation leads to perfect accuracy across all heuristic categories in the HANS test set. However, the second approach depends heavily on predefined patterns in the augmented data and may not fundamentally alter the model's learning process. Our study highlights the need for combining architectural innovations with data augmentation to enhance the robustness of NLI models.

1 Introduction

Natural Language Inference (NLI) tasks are used to evaluate a model's ability to understand and reason in natural language. Human-like NLI involves determining the logical relationships between a pair of sentences, and, despite the huge advancements of natural language models, it has become apparent that language models rely on heuristics, or superficial lexical cues, as opposed to learning the originally intended logical structures.

Research has shown that NLI models frequently exploit heuristics such as lexical overlap, subsequence and constituent (McCoy et al, 2019), or fail to handle simple lexical inferences involving synonyms and antonyms (Glockner et al, 2018). While these shortcuts may yield high accuracy on standard benchmarks, they falter when tested on datasets that are specifically designed to test it.

This paper aims to analyze the similar patterns observed in the Electra-small model (Clark et al., 2020), which was not in the original paper and tries to alleviate their limitations. The primary challenge lies in evaluating the extent of the model's dependency on these heuristics and proposing solutions that enable the model to generalize beyond simple pattern matching while also evaluating the lexical knowledge of a model. This work investigates how well the Electra small model performs under these conditions, providing insight into the broader issue of robustness in NLP. The structure of the report is as follows:

First, we evaluate the baseline performance of the Electra-small model on standard NLI datasets (MNLI and SNLI) to establish a reference for its performance and later examine its reliance on heuristics.

Next, in the error analysis part, we use adversarial datasets such as the test set requiring lexical knowledge in Glockner et al. (2018) and HANS (McCoy et al., 2019), we reveal how errors stem from heuristic dependencies. These analyses highlight where syntactic understanding should override shallow heuristic pattern-matching.

Finally, we explore potential solutions. We begin with an exploratory approach to architectural modification, adapting the decomposable attention model to Electra-small. This adaptation explicitly incorporates cross-sentence attention scores, aiming to encourage deeper reasoning and mitigate heuristic dependencies. Additionally, we investigate the effectiveness of data augmentation strategies, including fine-tuning with adversarial examples inspired by McCoy et al. (2019). These efforts provide valuable insights into improving robustness in NLI models while highlighting the challenges inherent in reducing reliance on shallow heuristics.

2 Datasets and Baseline

2.1 Data Sources

To comprehensively evaluate the performance of the Electra-small model and its reliance on syntactic heuristics, we utilize the following datasets:

SNLI (Stanford Natural Language Inference, Bowman et al. (2015)): A standard NLI benchmark dataset comprising entailment, contradiction, and neutral relations.

MNLI (Multi-Genre Natural Language Inference, Williams et al. (2018)): An extension of SNLI designed to cover diverse linguistic genres.

HANS (Heuristic Analysis for NLI Systems): An adversarial test set developed by McCoy et al. (2018) to evaluate NLI models' reliance on syntactic heuristics including lexical overlap, subsequence, and constituent. Details of the heuristics are provided in the next section.

GL: Adversarial test set from Glockner et al. (2018), used to evaluate robustness and generalization for lexical knowledge by creating minimal pairs that test a model's ability to recognize semantic entailment or contradiction. We will refer to this dataset as GL.

2.2 Baseline Model

To evaluate the performance of the Electra-small model, we conducted experiments under three configurations, and we selected Electra-small fine-tuned on a combination of the MNLI and SNLI datasets as our baseline model, sometimes referred to as the MNLI+SNLI model in the paper.

Table 1 presents the results under these different training configurations. The high accuracy of the baseline model on the SNLI (0.893) and MNLI (0.825) test sets demonstrates that it is well-tuned for these datasets.

Interestingly, Table 1 also shows that the model trained exclusively on the MNLI dataset performs well on the SNLI test set, suggesting that training on similar tasks can indeed yield transferable performance.

The accuracy of the listed models on the GL test set is also included in Table 1. A more detailed discussion of these results will be provided in the next section.

Model	Accuracy		
	SNLI	MNLI	GL
Electra (pre-trained)	0.352	0.322	0.710
Electra on MNLI	0.771	0.827	0.903
Electra on MNLI+SNLI (baseline)	0.893	0.825	0.948

Table 1: Accuracy results of the Electra-small models under different training configurations.

3 Error Analysis

3.1 Attack with GL (Glockner Test Set)

The Glockner test set is designed to assess a model's lexical knowledge by introducing

targeted adversarial examples. The errors are supposed to be apparent as in the original paper. However, as shown in Table 1, the pre-trained Electra-small model achieves a decent accuracy of 0.710 on this dataset, suggesting that the pre-trained embeddings already capture substantial lexical knowledge. When fine-tuned on MNLI or MNLI+SNLI, the Electra-small model demonstrates even higher accuracy (0.903 and 0.948), outperforming many models tested in the original study (Glockner et al., 2018).

Based on these results, we tentatively conclude that adversarial attacks using the Glockner test set are outdated, as more advanced models like Electra appear to exhibit stronger lexical knowledge, likely due to the large corpus of data in the pre-training phase. We now turn our focus to a more challenging dataset, HANS, in the following section.

3.2 Attack with HANS Test Set

The HANS dataset is specifically designed to expose and evaluate models' reliance on shallow syntactic heuristics (McCoy et al. 2019). McCoy et al. developed examples to test three specific heuristics:

Lexical overlap: Assumes that a model predicts entailment for any hypothesis with a high degree of overlap in words from the premise.

Subsequent: Assumes that a model predicts entailment for any hypothesis that is a contiguous subsequence of the premise. For example, the premise *"The doctor near the actor danced"* would entail the hypothesis *"The actor danced."*

Constituent: Assumes that a model predicts entailment for any hypothesis that forms a complete subtree of the premise's parse tree. For example, the premise *"If the artist slept, the actor ran"* would entail the hypothesis *"The artist slept."*

The primary goal of evaluating models on HANS is to assess their dependence on these shallow heuristics. As shown in Table 2, while the base model achieves near-perfect accuracy in heuristic-entailment cases, its low performance on non-entailment cases, particularly for the subsequent and constituent heuristics, highlights a dependency on syntactic shortcuts rather than deeper semantic reasoning.

Accuracy on Entailed Cases		
Lexical Overlap	Subsequence	Constituent
0.920	0.997	0.989
Accuracy on Non-Entailed Cases		
Lexical Overlap	Subsequence	Constituent
0.341	0.036	0.061

Table 2: Accuracy on the HASN test set for the baseline model (Electra on MNLI+SNLI)

Table 4 presents test examples originally provided by McCoy et al. (2018), adapted for our use. These specific examples further illustrate the baseline model's difficulty in distinguishing between entailment and non-entailment cases, as the model only predicts entailment for all examples, regardless of the correct classification.

To explore which words the baseline model emphasizes on, we examine a non-entailment lexical overlap example: the premise *"The judge by the actor stopped the banker"* and hypothesis *"The banker stopped the actor."* Figure 1a shows the saliency map for this example using the baseline model. The saliency map visualizes the estimated normalized gradient for each token, highlighting its contribution to the model's predictions. Figure 1a indicates that the highest weights are placed on the common words between the premise and hypothesis "*actor"* and "*banker"*, while the subject in "*judge"* in the premise receives a lower weight. This illustrates the model's reliance on overlapping words between the sentence pair.

4 Improving the Model: Attempts and Results

We attempt to improve the model's performance on HANS dataset first with architectural

modification, and second we investigate the effectiveness of data augmentation strategies.

4.1 Custom Decomposable Attention Model

Parikh et al. (2016) demonstrated that decomposing the NLI problem into pairwise comparisons between tokens or subphrases of the premise and hypothesis allowed their Decomposable Attention Model to outperform the best LSTM-based models of that time. Their results suggested that, for NLI tasks, pairwise comparisons might be more significant than global sentence-level representations.

Inspired by their work, we implemented a custom decomposable attention layer on top of the Electra-small model (Table 3). Specifically, we extracted the final-layer hidden states for the premise and hypothesis and computed pooled cross-attention scores for the sentence pair. These attention scores were then combined with the [CLS] token's hidden state and passed into a classifier.

Our rationale is that explicitly separating and comparing the contextualized representations of tokens in the premise and hypothesis may encourage the model to focus more on logical structures, rather than relying on superficial lexical overlaps.

The custom decomposable attention (custom DA) model was trained on the combined MNLI and SNLI datasets, achieving an accuracy of 0.860 on the combined test set—comparable to the baseline model trained on the same data (0.859). Notably, the custom DA model demonstrated significant improvement on the lexical overlap heuristic, achieving an accuracy of 0.664 compared to the baseline's 0.341 (Figure 2). However, the model continued to struggle with the subsequence and constituent heuristics, as illustrated by examples from McCoy et al. (2019) provided in Table 4. Encouragingly, the custom DA model successfully classified non-entailment cases in Table 4 for the lexical overlap heuristic. While the custom DA model did not effectively address the subsequence and constituent heuristics, its substantial improvement on the lexical overlap heuristic—although still below human-level accuracy—is promising. These results support our hypothesis that encouraging the model to focus more explicitly on pairwise comparisons between sentence pairs can lead to better performance on tasks involving logical relationships.

Table 3: Custom Decomposable Attention Layer

1. **Input Representation**: Extract final-layer hidden states for the [CLS] token and sentence pair (premium, hypothesis) using Electra-small.
2. **Cross-Attention Scoring**:
   a. Apply separate attention layers to premise and hypothesis hidden states.
   b. Compute cross-attention scores between premise and hypothesis.
   c. Pool attention scores for each premise token over all hypothesis tokens, and vice versa.
3. **Feature Combination**: Concatenate the [CLS] hidden state with pooled attention features.
4. **Classification**: Pass the combined features through a classifier to produce logits.

4.2 Data Augmentation

We explored data augmentation strategies utilized by McCoy et al. (2019). Specifically, we augmented the training data with adversarial examples designed to challenge the heuristics that the model appears to exploit. We augmented our training set by combining the original MNLI and SNLI datasets with the HANS training data. The HANS training set is specifically designed to challenge NLI models by including examples that exhibit features that will mislead models relying on shallow heuristics. By exposing the model to these challenging examples during training, we aimed to encourage it to develop a deeper understanding of the underlying semantics rather than relying on superficial

Heuristic	Premise	Hypothesis	Gold Label	MNLI + SNLI	Custom DA
Lexical overlap heuristic	The banker near the judge saw the actor.	The banker saw the actor.	E	E	E
	The lawyer was advised by the actor.	The actor advised the lawyer.	E	E	E
	The doctors visited the lawyer.	The lawyer visited the doctors.	N	E	N
	The judge by the actor stopped the banker.	The banker stopped the actor.	N	E	N
Subsequence heuristic	The artist and the student called the judge.	The student called the judge.	E	E	E
	Angry tourists helped the lawyer.	Tourists helped the lawyer.	E	E	E
	The judges heard the actors resigned.	The judges heard the actors.	N	E	E
	The senator near the lawyer danced.	The lawyer danced.	N	E	E
Constituent heuristic	Before the actor slept, the senator ran.	The actor slept.	E	E	N
	The lawyer knew that the judges shouted.	The judges shouted.	E	E	E
	If the actor slept, the judge saw the artist.	The actor slept.	N	E	E
	The lawyers resigned, or the artist slept.	The artist slept.	N	E	E

Table 4: Examples from McCoy et al. (2019) used to evaluate the three heuristics. Each sentence pair is accompanied by the gold label and the predicted labels from the baseline model (MNLI+SNLI) and the custom DA model. *E* represents entailment, and *N* represents non-entailment.

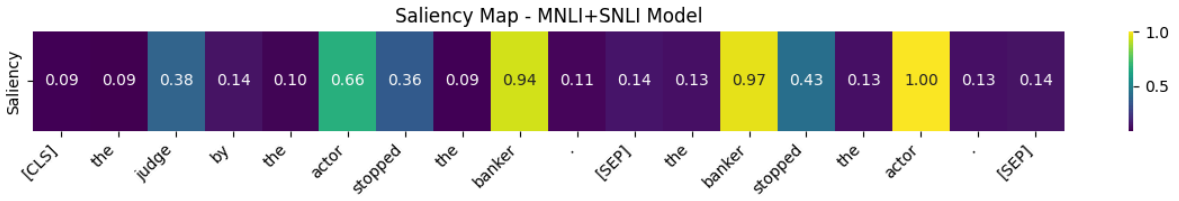
syntactic patterns.

After training the Electra-small model on the augmented dataset, we observed perfect improvements in performance on the HANS test set. As presented in Figure 2, the model achieved perfect accuracy across all heuristic categories in non-entailment cases, indicating a zero dependence on syntactic heuristics. When compared to the custom decomposable attention model (Custom DA) discussed in 4.1, data augmentation with the HANS training set provided substantial improvements across all heuristic categories.

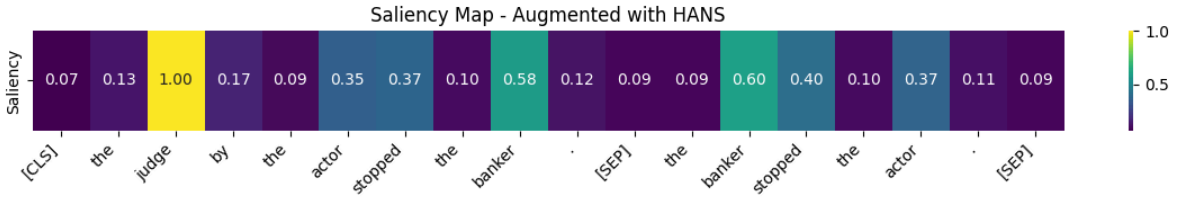
The custom decomposable attention model specifically improved performance on the lexical overlap heuristic but had limited impact on the subsequence and constituent heuristics. In contrast, the data augmentation approach completely mitigated heuristic reliance in non-entailment cases, as evidenced by the perfect accuracy scores. This mitigation is further supported by the saliency map in Figure 1b, where the highest weight is now placed on the critical term "*judge"*. While data augmentation provides perfect results on the HANS test set, it relies heavily on predefined patterns present in the augmented data. This approach may not fundamentally alter the model's underlying learning process but instead teaches it to recognize and avoid specific heuristic traps included in the training data. As a result, the model may remain vulnerable to other types of heuristics or adversarial examples not presented in the augmented dataset. These findings suggest that while data augmentation is a powerful tool for improving model robustness against known heuristic biases, it may need to be combined with architectural changes or other strategies to promote genuine understanding and generalization in NLI models.

5 Conclusions

This study explored the reliance of the Electra-small NLI model on syntactic heuristics and methods to mitigate this dependency. Using the HANS dataset, we found that while the model performs well on standard benchmarks like MNLI and SNLI, it heavily relies on superficial heuristics, leading to poor performance on non-entailment cases designed to exploit these shortcuts. We first implemented a custom decomposable attention model inspired by Parikh et al. (2016) to encourage pairwise



(a)



(b)

Figure 1: Saliency map for the example sentence pair, comparing (a) the baseline model (MNLI+SNLI) and (b) the model augmented with HANS.

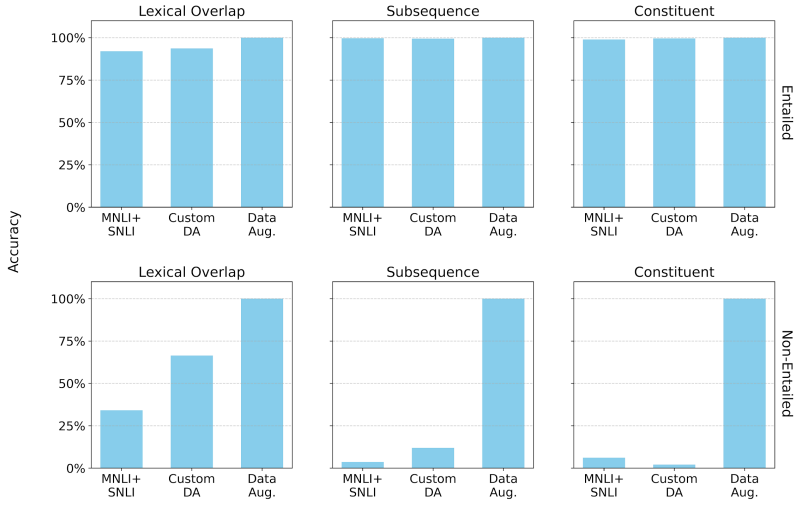


Figure 2: Accuracy on the HANS test set for the baseline model (MNLI+SNLI), the custom decomposable attention model (Custom DA) and the model augmented with HANS (Data Aug.).

comparisons between the premise and hypothesis. This modification improved accuracy on the lexical overlap heuristic from 0.341 to 0.664 but had limited impact on subsequence and constituent heuristics. Next, we incorporated the HANS training data for data augmentation. This approach proved highly effective, achieving perfect accuracy across all heuristic categories on the HANS test set and significantly reducing reliance on superficial

syntactic patterns. However, it depends heavily on the patterns in the augmented data and may not fundamentally change the model's learning process, leaving potential vulnerabilities to other heuristics not in the training data. Our findings suggest that while data augmentation is a powerful tool for improving model robustness against known heuristic biases, it may need to be combined with architectural changes or other strategies to promote genuine understanding and generalization in NLI models.. A combined approach that integrates architectural innovations alongside data augmentation could offer a more comprehensive solution to mitigate heuristic reliance while fostering deeper semantic comprehension.

References

[Bowman et al. 2015] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. Alarge annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

[Clark et al. 2020] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pretraining Text Encoders as Discriminators Rather Than Generators. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[Glockner et al. 2018] Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for 6 Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia, July. Association for Computational Linguistics.

[McCoy et al. 2019] Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy, July. Association for Computational Linguistics.

[Parikh et al. 2016] Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas, June. Association for Computational Linguistics.

[Williams et al. 2018] 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, New Orleans, Louisiana, June. Association for Computational Linguistics.