Towards preserving word order importance through FORCED INVALIDATION

Hadeel Al-Negheimish[1], **Pranava Madhyastha**[2,1] and **Alessandra Russo**[1]
[1] Imperial College London [2] City, University of London
{halnegheimish,a.russo}@imperial.ac.uk,
pranava.madhyastha@city.ac.uk

Abstract

Large pre-trained language models such as BERT have been widely used as a framework for natural language understanding (NLU) tasks. However, recent findings have revealed that pre-trained language models are insensitive to word order. The performance on NLU tasks remains unchanged even after randomly permuting the word of a sentence, where crucial syntactic information is destroyed. To help preserve the importance of word order, we propose a simple approach called FORCED INVALIDATION (FI): forcing the model to identify permuted sequences as invalid samples. We perform an extensive evaluation of our approach on various English NLU and QA based tasks over BERT-based and attention-based models over word embeddings. Our experiments demonstrate that FI significantly improves the sensitivity of the models to word order.[1]

1 Introduction

Ordering of words in a sentence is an important structural attribute for natural languages such as English, where subject-verb-object (SVO) structure is common and important to convey the meaning of the sentence. Understanding and comprehending natural language without strict adherence to a systematic ordering of words would make it an extremely challenging task. Recent work has investigated a surprising lack of sensitivity to word order information in state-of-the-art masked language models.

Recent research has focused on the impact of word order perturbation during the evaluation of models trained on well-ordered data. The results show that masked language models exhibit a catastrophic lack of sensitivity to word order permutations or shuffles, even for complex tasks in which task-relevant syntactic properties are completely destroyed (Pham et al., 2020; Al-Negheimish et al., 2021; Sinha et al., 2021b; Gupta et al., 2021). These studies show that models are still predicting the gold label for examples even after sequences have been permuted, and they also do so with high confidence (Sinha et al., 2021b; Gupta et al., 2021). This anomalous behaviour can potentially result in undesirable shortcuts or can cause models to fail catastrophically in simple adversarial settings. Furthermore, Sinha et al. (2021a) study the effect of pre-training masked language models on shuffled data, and suggest that the model might simply be capturing higher-order word co-occurrence statistics, rather than uncovering sophisticated semantic and syntactic structures necessary for language understanding.

In this paper, we present a simple, yet general approach called FORCED INVALIDATION (FI), where we force models to explicitly identify sequence permutations (§3). While our proposal is extensible to multiple types of models, here we present a controlled study over masked language model-based BERT models, and attention-based models over word-embeddings (§4). We present a large battery of experiments over a variety of natural language understanding tasks in the English language, including complex question answering based tasks, natural language inference based tasks and commonsense reasoning based tasks. Results show that our proposal significantly improves the sensitivity of the models to word-order information (§5).

2 Related Work

Recent work has proposed a few mitigation strategies for classification-type tasks: Pham et al. (2020) proposes improving word-order sensitivity by including a precursor fine-tuning step on synthetic CoLA-like tasks before finetuning for downstream tasks. While this improves their defined word order sensitivity score, accuracy on permuted samples remains significantly above chance, making this

[1] Our code and data for replication are available at https://github.com/halnegheimish/ForcedInvalidation

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2563–2570
May 2-6, 2023

approach unreliable. Gupta et al. (2021) present three approaches: one based on entropy regularisation, another on model probabilities thresholding, and finally an approach based on augmenting additional data consisting of destructive transformations, which include 1-gram permutations. The model is modified with an additional class to identify these destructive transformations. While the first two approaches require changes to model setup, the last one is based on a set of manual heuristics. Our approach has some similarity with the latter, however, our permutations are based on the principals borrowed from the n-gram language modelling literature (Roark et al., 2007), where the n-grams capture sufficient first-order statistics of the language. Our approach significantly reduces the models' reliance only on simple first-order statistics of language, and our empirical observations demonstrate the generalisability of our approach to various settings.

3 Methodology: FORCED INVALIDATION

Our method is grounded on recent observations which show that masked language models and other similar models tend to exploit shortcuts based on information about distributional word vicinity information for a diverse set of natural language processing tasks (Al-Negheimish et al., 2021; Sinha et al., 2021a). These shortcuts tend to make the models less sensitive to word-order information even for tasks that require the preservation of word ordering for accurate recovery of meaning (Sinha et al., 2021b). Al-Negheimish et al. (2021), in particular, show the insensitivity on a variety of n-gram based permutations of samples, where the authors specifically experiment with {1,2,3}-gram permutations. Further, {1,2,3}-gram based permutations capture a variety of word vicinity based patterns and also capture some of the most frequently occurring unigram, bigram and trigram patterns in a variety of benchmark datasets. Based on these salient observations, for each given dataset, the FORCED INVALIDATION (FI) methodology consists of the following two steps:

1. Augmenting training data with {1,2,3}-gram permutation samples (sampled from trainset) labelled with *invalid* as the additional label.

2. Modifying models to account for the new label and training them in the standard setting with a combination of standard training examples and the augmented *invalid* samples.

We observe that this simple FI approach improves the sensitivity of the model to word order and also improves the robustness of the models across a variety of tasks to first-order shortcuts. In the following sections, we present a rigorous experimental study that showcases the utility of FI.

4 Experimental settings

Data To generate n-gram permutations, we simply subsample from the training dataset, such that the ratio to the valid samples (samples with correct word order and the task-specific label) and invalid samples (samples with permuted n-grams and the *invalid* label) is 1-1. The invalid samples are generated such that they contain a uniform distribution of {1,2,3}-gram permutations.[2] Furthermore, we split the training set such that we use 90% of the samples for training the models and the remaining 10% is used as a development set. The development set is used to monitor training and perform early-stopping. Evaluations are done on a separate unseen task-specific validation set provided by the dataset creators.

We perform two evaluations: well-ordered and permuted. The first one is the standard evaluation of the model over the original unperturbed task-specific unseen validation set. For our experimental evaluations over permutations, we retain the original label of the same unseen validation set, but we permute the specific components of samples (e.g., premise or the hypothesis, question or the passage, etc.); we expect the models to reject the permuted sample instead of predicting the same ground-truth label. This setup allows us to evaluate the sensitivity of the model to word-order permutations of various degrees over the different components of the data.

Models We predominantly experiment with BERT-based models that either use BERT representations as the contextualised embeddings or classifiers that are directly trained with BERT. We also experiment with an additional simpler model that largely exploits attention over word-embeddings.[3] We will expand on the specific models for each of the tasks in the following section. Results for FI are

[2] Input string is divided to n-grams (based on white-space), and permuted, preserving the final punctuation. The only condition is that it varies from the original string, which is a weaker constraint than previous studies that require that no n-gram stays in its original position.

[3] Details about training parameters can be found in the appendix A

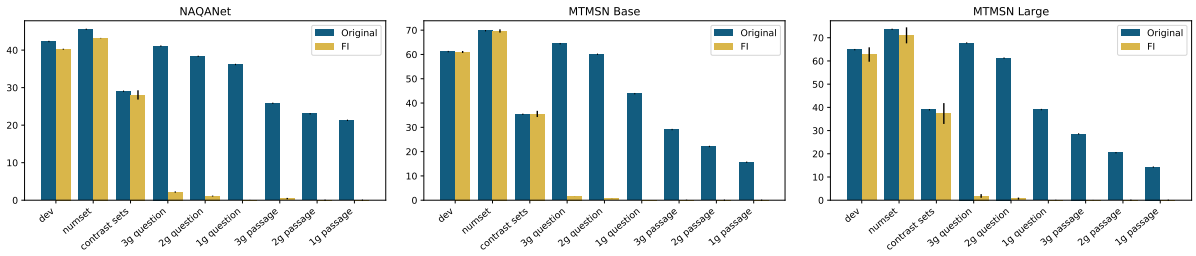

Figure 1: Accuracy using exact match (EM) on the DROP QA dataset, (a) NAQANet, (b) MTMSN Base, and (c) MTMSN Large. dev, numset and contrast sets are unperturbed well-ordered datasets, while the other bars show {1,2,3}-gram permutations on numset. FI models exhibit clear sensitivity to word order, as they no longer predict the original answer in most perturbed cases. FI accuracy is averaged over training with 3-random seeds, and the standard deviation is shown over the yellow bars. FI-models are confident in identifying invalid samples.

reported over the averages and standard deviation of models trained with three random seeds.

5 Results and Observations

5.1 Unconstrained Question Answering

DROP (Dua et al., 2019) is a reading-comprehension dataset with unconstrained answers. It comprises questions that require reasoning over the content of different paragraphs. Even though this is designed to be a challenging task, Al-Negheimish et al. (2021) show that for most models, permuting questions had little impact on the model's ability to predict the correct answer for numerical reasoning questions. This was found specifically problematic as the task consists of complex questions, and permutations render the questions syntactically and semantically redundant. We apply FI for two module-based models designed specifically for this task, NAQANet (Dua et al., 2019), an attention-based model with GloVe-based embeddings (Pennington et al., 2014), which is the original model proposed by the authors of the dataset; and MTMSN (Hu et al., 2019), which is based on BERT-based (Devlin et al., 2019) contextual representations. These models have separate modules targeting different kinds of reasoning, e.g. a module for counting and a module for arithmetic expressions. We augment these with an additional module to force invalidation over invalid permuted samples, a two-way classification module that learns to distinguish between permuting either the question or passage. FI models should now choose the invalid type for permuted samples, instead of giving the same answer as well-ordered samples. Fig 1 shows a comparison of the Exact Match accuracy of NAQANet and MTMSN, between original and FI models. We observe that question

			Part 1			Part 2		
	Variation	dev	3-gram	2-gram	1-gram	3-gram	2-gram	1-gram
UQA	NAQANet	3.15	94.64	97.18	99.59	99.01	99.93	100.00
	MTMSN Base	0.19	97.14	98.99	99.91	99.80	99.99	100.00
	MTMSN Large	0.07	95.01	97.52	98.98	99.43	100.00	100.00
NLI	RTE	0.36	94.57	98.19	99.64	96.01	96.74	98.55
	MNLI_M	1.36	94.52	97.60	99.21	94.83	97.35	99.45
	MNLI_MM	1.25	96.20	98.17	99.49	94.47	97.43	99.51
	ANLI1	0.70	99.80	100.00	100.00	95.20	98.10	99.69
	ANLI2	0.50	99.90	100.00	100.00	93.89	98.50	99.10
	ANLI3	1.33	99.83	99.92	100.00	94.83	97.33	98.67
GA	CoLA	2.30	92.66	92.35	97.93	-	-	-

Table 1: Percentage of evaluation data predicted as invalid in FI models in all of the tasks. dev is the unperturbed validation set. {n}-gram permutations of part 1 correspond to permutations of the question in UQA and of the premise in NLI. CoLA is made up of single sentences. All models succeed at flagging these samples as invalid.

permutations have little effect on the original model, as previously noted in (Al-Negheimish et al., 2021). Interestingly, passage permutations can drastically reduce performance by a third. While performance degrades for passage permutations, it remains unacceptably high, as we note that DROP is an unconstrained-QA task, so the space of possible answers is large. FI, on the other hand, succeeds in making the model sensitive to almost all permutations. We see that this generalises across $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$. More importantly, we observe that while the model with FI is sensitive to word order (no longer predicting original answers for permuted examples), the model's performance on well-ordered data is largely retained. To demonstrate that FI are correctly predicting invalid, table 1 shows the percentage of the data predicted as invalid, which is near-perfect for permuted samples.

5.2 Grammatical Acceptability

CoLA (Wang et al., 2018) is a task that measures models' ability to determine the grammatical acceptability of sentences. Pham et al. (2020) show that from the GLUE benchmark (Wang et al., 2018),

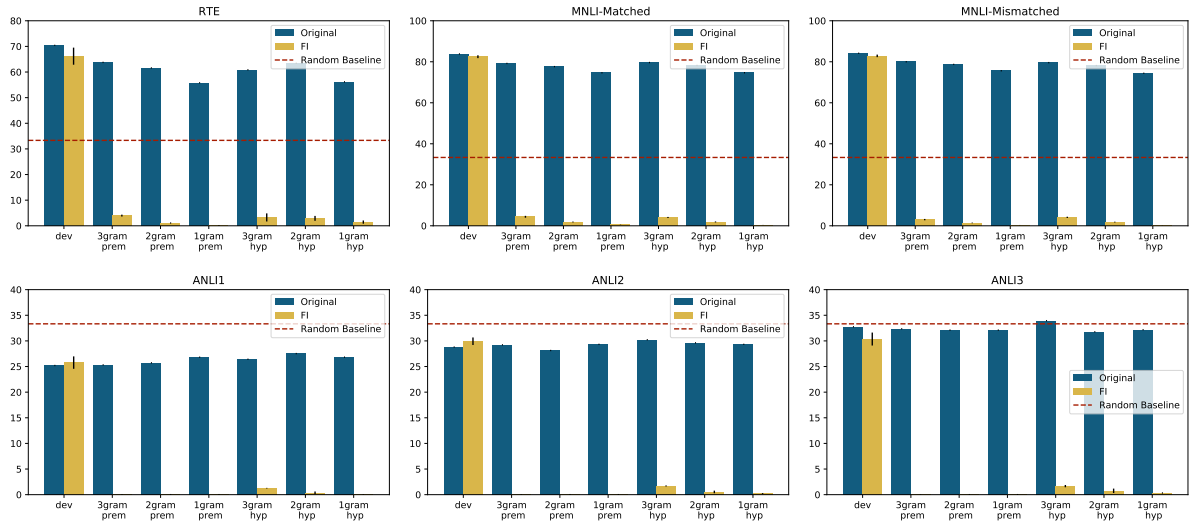

Figure 2: Accuracy for NLI tasks (a) trained on RTE, (b-f) trained on MNLI data. Original models exhibit a lack of sensitivity to word order, as they have the same accuracy regardless of the n-gram permutations. FI models are able to tell apart invalid examples even in out-of-distribution ANLI data.

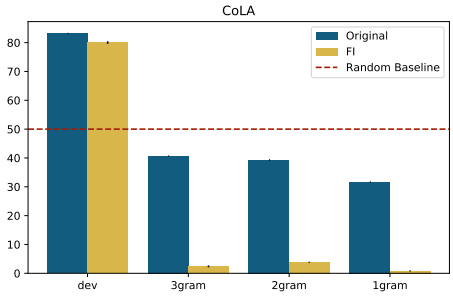

Figure 3: Accuracy of original models and FI on CoLA grammatical acceptability task. While the original model is below chance, we note that because this task is grammar-detection, the original model should not accept any of the permuted samples.

this task requires models to be most sensitive to word order. We applied FI by extending the BERT-based classification model with another '*invalid*' class label, to flag permuted sentences. Grammar and syntax have been destroyed in permuted sentences, we would expect the original model to label them as *not acceptable*. However, we see in Fig 3 that it maintains significantly high accuracy for 3-gram and 2-gram permutations (note that CoLA is an imbalanced dataset (70% acceptable)). Concretely, we observe that the standard BERT-based model labels $185/967$ 3-gram permutations, and $114/967$ 2-gram permutations as *acceptable*. Our FI approach significantly ameliorates this problem by increasing the sensitivity of the model to word-order permutations, as we observe in Fig 3.

5.3 Natural Language Inference (NLI)

NLI has been one of the important testbeds of previous work studying BERT-based models and their lack of sensitivity to word-order information (Sinha et al., 2021b; Abdou et al., 2022). These studies suggest that BERT-based models almost always assign the same labels to examples with perturbed word order as well-ordered ones highlighting the lack of sensitivity to word order, and likely dependence on shallow features. Similarly to §5.2, applying FI to an NLI BERT model was done by simply extending it with another class to represent *invalid* input. We perform a battery of experiments over a variety of NLI tasks in GLUE (Wang et al., 2018) such as RTE and MNLI (Williams et al., 2018) tasks. FI makes models highly sensitive to permuted sequences, as shown in Fig. 2. We verify in table 1 that those invalid sentences were correctly flagged as invalid. We also observe that FI-based models trained on MNLI and tested on out-of-distribution ANLI (Nie et al., 2020) show extreme sensitivity to word-order perturbations; this shows that the model has learned to flag invalid input and generalise to similar unseen tasks. Additionally, we observe that FI makes the models less likely to suffer from shortcut effects, which are common in models trained for NLI tasks (McCoy et al., 2019). Further experiments on Arabic NLI are presented in Appendix D, where we demonstrate that FI works well for other languages beyond English.

Heuristic Analysis

McCoy et al. (2019) introduces an evaluation set *Heuristic Analysis for NLI Systems (HANS)*, that examines surface heuristics NLI models are prone to adopting. They show in the existence

		Lexical Overlap	Subsequence	Constituent
Original	Entailment	97.76	99.92	100.00
	Non-Entailment	14.74	0.58	2.66
FI	Entailment	80.94	99.56	99.60
	Non-Entailment	64.56	38.14	11.48

Table 2: Comparison of BERT finetuned on MNLI with and without FI, FI makes the model more robust to the syntactic heuristics presented in (McCoy et al., 2019)

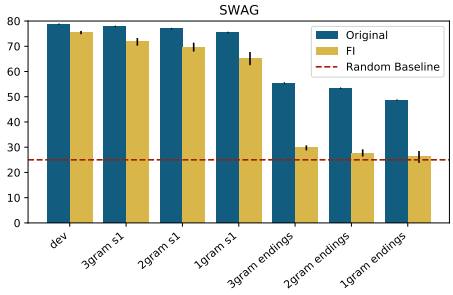

Figure 4: Accuracy of original models and FI on SWAG multiple-choice commonsense reasoning task, both are presented with the same four answer choices, FI enforces sensitivity to word order, especially on the endings.

of these heuristics, models always predict entailment, as they have near-perfect accuracy for that label, but perform poorly when the actual label is non-entailment (accuracy close to 0% in most cases). The heuristics targeted by this dataset are special cases of each other: the most general, *lexical overlap* heuristic: assume that all hypotheses constructed from words in the premise are entailed. Next follows is the *subsequence* heuristic: assume all hypotheses made up of contiguous subsequences of the premise are entailed. Finally, the *constituent* heuristic: assume all hypotheses made up of complete subtrees of the premise's parse tree are entailed. We expect that FI-models will be more robust to these shortcuts, and find (table 2) that this is indeed the case for lexical-overlap and the more challenging subsequence heuristics, substantially improving non-entailment performance, indicating that the models are no longer strictly biased with the presence of these heuristics. We note a surprising result of a drop in accuracy for entailed lexical overlap samples, which could be caused by the model no longer taking that shortcut, and warrants additional investigation in future work.

5.4 Multiple Choice Commonsense Reasoning

SWAG (Zellers et al., 2018) is a commonsense and grounded reasoning task that requires choosing between different possible ending scenarios given some context, where the context comprises of a primary sentence followed by the initial set of words for the following sentence. The model is trained to predict the most likely ending scenario given context and a list of four ending scenarios. To train our FORCED INVALIDATION model, we add an additional answer choice `'is invalid.'` to the data and perform n-gram based permutation of the primary sentence or over each of the endings. Nothing is changed in the architecture of the model. For evaluation, we maintain the same datasets for the Original and FI models, so they only contain the same number of answer choices, without the `'invalid'` choice. Ideally, the models should achieve random performance in the permuted cases. We present our results in Fig. 4; where firstly we observe that FI does not seem to reliably affect the model's sensitivity to word order perturbations with all the combinations. We specifically observe that n-gram permutations of primary sentence have little impact on the performance of FI-models and they seem to be less sensitive to word-order perturbations than expected. However, n-gram permutations of endings result in FI-models obtaining near-random performance as expected. We investigated the cause of this anomaly and observed that SWAG dataset has a prominent problem, in that, the primary sentence for a majority of cases is almost irrelevant to the model. A model trained on SWAG dataset is able to predict the correct answer for 60% of the examples without having access to the primary sentence.

6 Conclusion

In this paper, we presented a simple yet general technique called FORCED INVALIDATION, that significantly improves the sensitivity of models towards word order information. Our methodology requires minimal changes to the model and is sample efficient and drastically increases the sensitivity of the models to permutations of word order for a variety of tasks. We present a focused empirical validation of our methodology to showcase its generalisability. While in this paper we have focused on masked language models and attention-based models over word embeddings, we expect FI to generalise for other modelling setups such as RNN-based and CNN-based models and leave it as future work. We anticipate that this approach will also serve as a solution for other undesired behaviors in the model by explicitly invalidating such behaviours. We leave this as future work.

Limitations

While our empirical results showcase the effectiveness of FI and increase models' sensitivity to word order, the causal mechanisms are not currently obvious. It is not clear whether or not positional encodings are reflecting this change. Like previous work, our observations are additionally only restricted currently to English and Arabic (appendix D), further experiments are required to establish the problems relating to word order sensitivity and the utility of FI for other languages.

Acknowledgements

This research has been supported by a PhD scholarship from King Saud University. We thank our anonymous reviewers for their constructive feedback.

References

Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev, and Anders Søgaard. 2022. Word order does matter and shuffled language models know it. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6907–6919, Dublin, Ireland. Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Hadeel Al-Negheimish, Pranava Madhyastha, and Alessandra Russo. 2021. Numerical reasoning in machine reading comprehension tasks: are we there yet? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9643–9649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

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.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. of NAACL*.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. 2021. BERT & family eat word salad: Experiments with text understanding. *CoRR*, abs/2101.03453.

Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. 2019. A multi-type multi-span network for reading comprehension that requires discrete reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1596–1606, Hong Kong, China. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

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, Doha, Qatar. Association for Computational Linguistics.

Thang M. Pham, Trung Bui, Long Mai, and Anh Nguyen. 2020. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? *CoRR*, abs/2012.15180.

Brian Roark, Murat Saraclar, and Michael Collins. 2007. Discriminative n-gram language modeling. *Computer Speech & Language*, 21(2):373–392.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021a. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021b. UnNatural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7329–7346, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. 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, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

A Training Details

We use $\text{BERT}_{\text{BASE}}$ models with a classification head on top (BERTForSequenceClassification (Wolf et al., 2020)) for NLI and CoLA. It was trained for 5 epochs, batch size 16, learning rate 2e-5. We use a $\text{BERT}_{\text{BASE}}$ model (Wolf et al., 2020), for SWAG. It was trained for 3 epochs, batch size 16, learning rate 5e-5. All of the above are done using google colab with high-ram. MTMSN is based on the published codebase (Hu et al., 2019), and we use the same parameters to train, namely: BERT_{base}: batch size 24, 5 epochs and learning rate 3e-5, BERT_{large}: batch size 12, 10 epochs and learning rate 3e-5. Training was done on four RTX6000 GPUs with 24GB of RAM each for $\text{BERT}_{\text{LARGE}}$, and a single one was used for $\text{BERT}_{\text{BASE}}$. Models are trained for both settings, original and FI, to provide a fair comparison.

B Dataset Statistics

As described before, we filter out examples with sentences containing less than three words, such that we can generate at least 1 3-gram shuffle. Table 3 describes the tasks' validation sets before and after filtration.

	Original	Used	p1 #words	p2 #words
DROP numset	6848	6848	11	182
RTE	277	276	31	8
MNLI	9815	9289	16	9
MNLI-mm	9832	9551	17	10
ANLI1	1000	999	54	10
ANLI2	1000	999	54	9
ANLI3	1200	1199	52	9
CoLA	1043	967	7	-
SWAG	20006	19352	11	8

Table 3: Statistics of the validation set of datasets used for evaluation, including the original number of examples, after filtration, and median number of words for the first and second components.

Dataset licenses are mentioned in table 4:

	License
DROP numset	CC BY 4.0
RTE	Unknown
MNLI	OANC
MNLI-mm	OANC
ANLI1	CC BY-NC 4.0
ANLI2	CC BY-NC 4.0
ANLI3	CC BY-NC 4.0
CoLA	Unknown
SWAG	MIT license

Table 4: Artifact licenses for the datasets used.

C FI as a precursor to downstream task finetuning

Inspired by (Pham et al., 2020), we first perform FI on BERT to solely categorise valid and invalid samples for sentences that are sampled from Wikipedia. We replace the standard BERT-based contextual representations in MTMSN with FI-BERT based contextual embeddings, to see if it helps it become more sensitive to word order in the downstream DROP task. This did not show an improvement,

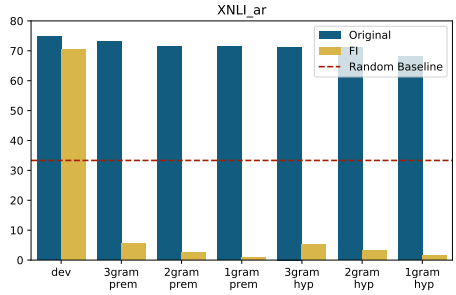

Figure 5: Accuracy of original models and FI on an Arabic NLI task. Once more, we see that the original models are insensitive to word order even in Arabic, while FI models learn to flag invalid samples.

however, where permuted examples are still predicted the same as well-ordered ones. This indicates that having an explicit way to flag invalid examples is helpful to the models.

D FORCED INVALIDATION with Other Languages

To establish the generalizability of this approach to other languages, we applied FIon an Arabic NLI task. We used the Arabic split of the XNLI dataset (Conneau et al., 2018) to finetune the Ar-BERT model (Abdul-Mageed et al., 2021), and compare the original training setup with FORCED INVALIDATION. Figure 5 shows that this approach successfully preserves the importance of word order beyond the English language.