Supervising Model Attention with Human Explanations for Robust Natural Language Inference

Joe Stacey[1], **Yonatan Belinkov**[2], **Marek Rei**[1]
[1]Imperial College London
[2]Technion – Israel Institute of Technology
j.stacey20@imperial.ac.uk, belinkov@technion.ac.il, marek.rei@imperial.ac.uk

Abstract

Natural Language Inference (NLI) models are known to learn from biases and artefacts within their training data, impacting how well they generalise to other unseen datasets. Existing de-biasing approaches focus on preventing the models from learning these biases, which can result in restrictive models and lower performance. We instead investigate teaching the model how a human would approach the NLI task, in order to learn features that will generalise better to previously unseen examples. Using natural language explanations, we supervise the model's attention weights to encourage more attention to be paid to the words present in the explanations, significantly improving model performance. Our experiments show that the in-distribution improvements of this method are also accompanied by out-of-distribution improvements, with the supervised models learning from features that generalise better to other NLI datasets. Analysis of the model indicates that human explanations encourage increased attention on the important words, with more attention paid to words in the premise and less attention paid to punctuation and stopwords.

Introduction

Natural Language Inference (NLI) models predict the relationship between a premise and hypothesis pair, deciding whether the hypothesis is entailed by the premise, contradicts the premise, or is neutral with respect to the premise. While NLI models achieve impressive in-distribution performance, they are known to learn from dataset-specific artefacts, impacting how well these models generalise on out-of-distribution examples (Gururangan et al. 2018; Tsuchiya 2018; Poliak et al. 2018). De-biasing efforts to date have successfully improved out-of-distribution results, but mostly at the expense of in-distribution performance (Belinkov et al. 2019a; Mahabadi, Belinkov, and Henderson 2020; Sanh et al. 2020).

While most previous work creating more robust NLI models has focused on preventing models learning from biases or artefacts in their datasets (more details in the Related Work section), we take a different approach. We aim to use information about how humans approach the task, training with natural language explanations in the e-SNLI dataset (Camburu et al. 2018) to create more robust models.

Premise:
*Wet brown **dog swims** towards camera.*

Hypothesis:
*A **dog** is **sleeping** in his bed.*

Explanation for contradiction class:
*A **dog** cannot be **sleeping** while he **swims**.*

Figure 1: An example of using a free text explanation to identify important words in the premise and hypothesis. In this case the words *dog*, *sleeping* and *swims* have been identified from the explanation.

Human explanations have been found to improve performance on a range of tasks (Rajani et al. 2019; Andreas, Klein, and Levine 2018; Mu, Liang, and Goodman 2020; Liang, Zou, and Yu 2020); however, this has largely not been the case in NLI (Hase and Bansal 2021; Kumar and Talukdar 2020; Camburu et al. 2018). Generating human explanations from e-SNLI has been found to improve model performance (Zhao and Vydiswaran 2021), but this process is highly computationally expensive and the in-distribution improvements are accompanied by a reduction in out-of-distribution performance. We aim to address both issues, proposing a simple and efficient method for using explanations to improve model robustness while also improving in-distribution performance.

We investigate multiple approaches to incorporate these human explanations. Firstly, we introduce an additional loss term to encourage the model to pay more attention to words in the explanation, supervising the attention from the [CLS] token in the existing model self-attention layers. Additionally, we introduce another attention layer on top of the model and supervise its weights. We also adapt a further attention-based approach for incorporating explanations as proposed by Pruthi et al. (2020), testing whether this method also improves performance and model robustness for NLI. Each approach considers the most important words in the hypothesis and premise based on the e-SNLI human explanations (see Figure 1).

To summarise our contributions: 1) We propose a method for supervising with human explanations that provides sig-

nificant improvements on both in-distribution and out-of-distribution NLI datasets. 2) We show that when combined with DeBERTa (He et al. 2021), this approach achieves a new state-of-the-art result for SNLI (Bowman et al. 2015). 3) We show that the model attention weights can effectively predict which words will appear in the explanations, reaching the same performance as prior work that focuses on this task. 4) Finally, we show that training with human explanations encourages the model to pay more attention to important words in the premise and focus less on stop-words in the hypothesis, helping to mitigate the hypothesis-only bias of NLI systems (Gururangan et al. 2018).[1]

Related Work

Training NLI Models with Explanations

Most work to date has found that training with NLI explanations does not translate into either in-distribution or out-of-distribution improvements (Camburu et al. 2018; Kumar and Talukdar 2020; Hase and Bansal 2021). Camburu et al. (2018) implement two approaches for incorporating the model explanations: using an *Explain then Predict* approach which generates an explanation and uses it to predict the class, and also predicting both the NLI class and generating the explanation from the same vector of features. Neither of these approaches significantly improved performance in-distribution or out-of-distribution on the MNLI dataset.

Hase and Bansal (2021) use a retrieval-based approach for incorporating the e-SNLI explanations, retrieving the top explanations for a hypothesis and premise pair and combining the sentences with the retrieved explanations. They conclude that the e-SNLI dataset does not meet the six preconditions for their retrieval approach to improve performance, with these conditions including how explanations need to be sufficiently relevant across data points.

Kumar and Talukdar (2020) generate explanations specific to each class, using these explanations along with the premise and hypothesis to predict the NLI class. This corresponds to a drop in performance both in-distribution and out-of-distribution (Kumar and Talukdar 2020). Zhao and Vydiswaran (2021) also generate explanations for each class, first predicting which of the words in a hypothesis are relevant given the class, training with the highlighted words in e-SNLI. Explanations are then generated based on these annotated hypotheses. While this approach did improve in-distribution performance, out-of-distribution performance did not improve. This process involved training a pipeline of three RoBERTa (Liu et al. 2019) models and a GPT2 (Radford et al. 2019) model, with the performance of this pipeline compared to the performance of a single RoBERTa baseline model.

Unlike the prior work, we aim to show how training with human explanations can improve out-of-distribution performance. We also aim to show that in-distribution improvements are possible within a single model, without requiring a pipeline of models, and that these in-distribution and out-of-distribution benefits can be achieved simultaneously.

[1]https://github.com/joestacey/NLI_with_a_human_touch

Training with Explanations Beyond NLI

Pruthi et al. (2020) introduce a teacher-student framework for training with explanations, finding that attention-based approaches are the most effective way to improve performance on sentiment analysis and question answering tasks. For sentiment analysis this involved supervising the attention from the [CLS] token. Attention-based methods to incorporate explanations have also been found to improve performance on hate speech detection (Mathew et al. 2021).

Closest to our work, Pruthi et al. (2020) supervise the average attention weights across all of a model's attention heads, whereas we identify which specific heads benefit the most from the supervision and then supervise these heads individually. Their method uses KL-Divergence as an auxiliary loss, while we found mean squared error to perform better when supervising attention. Moreover, Pruthi et al. (2020) do not consider out-of-distribution performance, which is the focus of our work, and do not use free-text explanations, while we incorporate explanations either as free-text explanations or in the form of highlighted words.

Pruthi et al. (2020) train with up to 1,200 and 2,500 examples across two tasks, while we train with a large corpus of 550,152 training observations. As there is more benefit from the explanations when training with fewer examples (Pruthi et al. 2020), it is also not clear whether the improvements will translate to a dataset of this scale. Pruthi et al. (2020) also investigate training with explanations for sentiment analysis and question answering tasks, whereas we train with explanations for NLI, a task where most prior work finds that explanations do not improve performance (Hase and Bansal 2021; Kumar and Talukdar 2020; Camburu et al. 2018). We investigate the performance from adapting the method proposed by Pruthi et al. (2020) to NLI, in addition to comparing this with the improvements from our two proposed approaches.

More widely, explanations have improved performance on a range of domains, including commonsense reasoning (Rajani et al. 2019), relation extraction (Murty, Koh, and Liang 2020) and visual classification tasks (Liang, Zou, and Yu 2020; Mu, Liang, and Goodman 2020). Prior work focuses on finding in-distribution improvements rather than considering model robustness, whereas we find that the largest impact from training with model explanations can be the corresponding improvements in model robustness.

Creating More Robust NLI Models

Previous work on creating more robust NLI models has focused on preventing models learning from artefacts (or *biases*) in their training data. The most common strategy for mitigating biases within NLI is by creating a weak model to intentionally learn a bias, then encouraging a target model to have low similarity to this weak model (He, Zha, and Wang 2019; Clark, Yatskar, and Zettlemoyer 2019; Mahabadi, Belinkov, and Henderson 2020; Utama, Moosavi, and Gurevych 2020b; Sanh et al. 2020; Liu et al. 2020; Clark, Yatskar, and Zettlemoyer 2020) or to use the weak model to weight training observations (Clark, Yatskar, and Zettlemoyer 2019; Utama, Moosavi, and Gurevych 2020b; Liu et al. 2020).

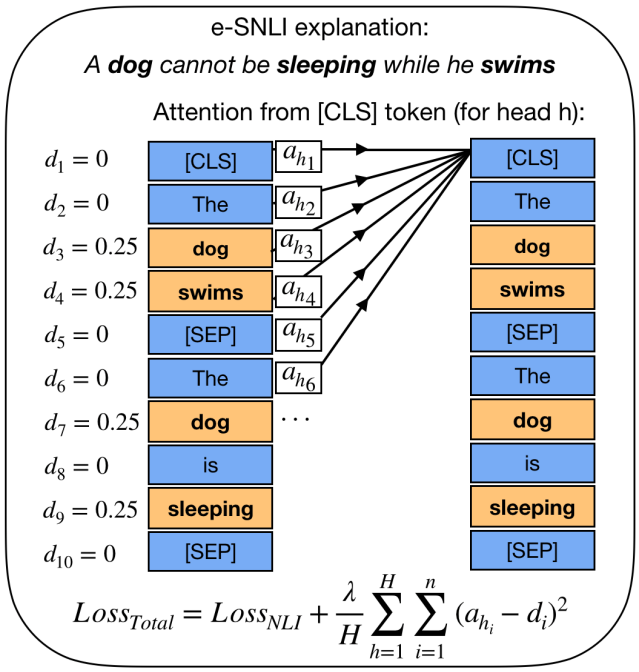

Figure 2: An example of how the attention loss is calculated when supervising an existing self-attention layer.

Other strategies to prevent models learning from artefacts include using adversarial training with gradient reversal to mitigate the hypothesis-only bias (Belinkov et al. 2019a,b; Stacey et al. 2020), using data-augmentation (Min et al. 2020; Minervini and Riedel 2018), fine-tuning on minority examples (Yaghoobzadeh et al. 2021), gradient supervision with counterfactual examples (Teney, Abbasnedjad, and van den Hengel 2020), multi-task learning (Tu et al. 2020) or creating compressed representations to remove irrelevant information (Mahabadi, Belinkov, and Henderson 2021). We take a new and different approach, encouraging models to learn from how humans would approach the task.

Attention Supervision Method

The e-SNLI explanations (Camburu et al. 2018) were created by asking Amazon Mechanical Turk annotators why each hypothesis and premise had their given label. The explanations take the form of either free text explanations, or highlighted words in the premise and hypothesis that annotators believe are important. Based on these explanations we create labels $E = \{e_i\}_{i=1}^{n}$ for each observation, with e_i taking values of either 0 or 1 to indicate whether a token is relevant to a human explanation, and n being the number of tokens in the NLI sentence pair. For free-text explanations, e_i has a value of 1 if its corresponding token is from a word present in the explanation, otherwise the value is 0. For the highlighted words, e_i has a value of 1 if the corresponding word in the premise or hypothesis has been highlighted by the annotator. For the free text explanations we exclude stop-words, whereas highlighted stopwords are selected.[2]

These explanations are only used during training, whereas during testing the model predicts the NLI class based on the hypothesis and premise alone.

Supervising Self-Attention Layers

To supervise the model's attention weights we create a desired distribution $D = \{d_i\}_{i=1}^{n}$ of attention values, normalizing the e_i values to sum to 1:

$$d_i = \frac{e_i}{\sum_{k=1}^{n} e_k}$$

We supervise the [CLS] attention weights in the final self-attention layer of a transformer model, introducing a second loss term to encourage assigning more attention to words in the human-annotated explanations (see Figure 2). We supervise the attention weights in the final self-attention layer as we find this performs better than supervising previous layers. Where a_{h_i} denotes the attention weights for a given attention head, the total loss is defined as:

$$Loss_{Total} = Loss_{NLI} + \frac{\lambda}{H}\sum_{h=1}^{H}\sum_{i=1}^{n}(a_{h_i} - d_i)^2$$

where $Loss_{NLI}$ is the main cross-entropy loss for the NLI task, H is the number of heads being supervised and λ is a hyper-parameter weighting the attention component of the model loss. The attention values for a given head a_{h_i} are defined as:

$$a_{h_i} = \frac{\exp\left(q_{h_{CLS}}^{T} k_{h_i} / \sqrt{d_k}\right)}{\sum_{j=1}^{n} \exp\left(q_{h_{CLS}}^{T} k_{h_j} / \sqrt{d_k}\right)}$$

Where $q_{h_{CLS}}$ represents the CLS query vector for the head, k_{h_i} are the key vectors for the other tokens in the sentence and d_k is the dimensionality of the key vectors.

Selecting Attention Heads for Supervision

As the attention heads can have different roles (Clark et al. 2019; Vig and Belinkov 2019), when supervising an existing self-attention layer we investigate how many and which heads should be supervised. We supervise each attention head in turn to investigate which heads benefit the most from the supervision. We then choose the top *K* heads for supervision, where *K* is a hyper-parameter tuned across the values $\{1, 3, 6, 9, 12\}$ using 5 random seeds for each condition. This greedy approach does not guarantee finding the optimal subset of heads, but it is more efficient than trying all subsets. By introducing this approach to selectively supervise the attention heads, the model can benefit from the explanation supervision while also allowing for diversity between the roles of the supervised and unsupervised attention heads.

[2]Performing the matching based on free text would return many incorrect stop-words, whereas using the highlights allows us to focus specifically on the ones that the annotators have selected.

	Dev	Test	Hard	MNLI mi	MNLI ma	ANLI	HANS
BERT baseline	90.05	89.77	79.36	72.52	72.28	31.81	56.83
Ours (extra layer)	90.40	90.09	79.96	73.03	73.10	31.47	57.85
Improvement	**+0.35**†‡	**+0.32**†‡	**+0.60**†‡	**+0.51**†	**+0.82**†‡	-0.34	+1.02
Ours (existing attention)	90.45	90.17	80.15	73.36	73.19	31.41	58.42
Improvement	**+0.40**†‡	**+0.40**†‡	**+0.79**†‡	**+0.84**†‡	**+0.91**†‡	-0.40	**+1.59** †

Table 1: Average accuracy across 25 random seeds, evaluated on: SNLI-dev, SNLI-test, SNLI-hard, MNLI mismatched (MNLI mi), MNLI matched (MNLI ma), ANLI (R1, R2 and R3) and HANS. Ours (extra layer) involves creating and supervising an additional attention layer on top of the model, while Ours (existing attention) involves supervising 3 heads of an existing self-attention layer. Significant results with P-values less than 0.05 are shown in bold and with a †. ‡ indicates results that are statistically significant after applying a Bonferroni correction factor of 7 for each dataset tested.

Supervising an Additional Attention Layer

Instead of supervising an existing self-attention layer in the model, an additional attention layer can also be created using the sequence representations $\{h_i\}$ from the transformer model. Using an architecture similar to Rei and Søgaard (2019), we define our unnormalised attention values $\widetilde{a_i}$ as:

$$\widetilde{a_i} = \sigma(W_{h2}(\tanh\left(W_{h1}h_i + b_{h1}\right)) + b_{h2})$$

where W_{h1} and W_{h2} are trainable parameters along with their respective bias terms. We supervise the normalized attention weights a_i:

$$a_i = \frac{\widetilde{a}_i}{\sum_{k=1}^{n} \widetilde{a}_k}$$

These weights are used to create a new representation c:

$$c = \sum_{i=1}^{n} a_i h_i$$

Finally, a linear classifier and softmax are applied to this representation to predict the class. $Loss_{Total}$ is the same as described previously, using the single attention head.

Experimental Setup and Evaluation

The attention supervision was implemented with BERT (Devlin et al. 2019) and DeBERTa (He et al. 2021), the latter using disentangled matrices on content and position vectors to compute the attention weights. We use DeBERTa to assess whether our proposed approach can improve on current state of the art results. λ was chosen based on performance on the validation set, trying values in the range $[0.2, 1.8]$ at increments of 0.2. For our BERT model the best performing λ is 1.0, equally weighting the two loss terms, whereas for DeBERTa this value was 0.8.

The robustness of the model is assessed by significance testing on the MultiNLI matched and mismatched validation sets (Williams, Nangia, and Bowman 2018), and the ANLI (Nie et al. 2020), SNLI-hard (Gururangan et al. 2018) and HANS (McCoy, Pavlick, and Linzen 2019) challenge sets, using a two-tailed t-test to assess significant improvements from the baseline. HANS contains examples where common syntactic heuristics fail, while SNLI-hard is created from the SNLI test set with examples that a hypothesis-only model has misclassified. ANLI is created using a human-in-the-loop setup to create intentionally challenging examples. The SNLI dev and test set are considered in-distribution, while HANS, ANLI, SNLI-hard and the MNLI mismatched and matched datasets are considered out-of-distribution.

Experiments

Performance in and out of Distribution

The experiments show that supervising the attention patterns of BERT based on human explanations simultaneously improves both in-distribution and out-of-distribution NLI performance (Table 1). When supervising an existing self-attention layer, in-distribution accuracy on the SNLI test set improves by 0.4%. The hard subset of this set, SNLI-hard, has a larger improvement of 0.79%, showing that the human explanations provide the most benefit for the hardest SNLI examples. The improvements in SNLI-test and SNLI-hard are significant, with p-values less than 10^{-8}. Moreover, out-of-distribution performance improves on both of the MNLI validation sets and on HANS, with accuracy improvements of 0.84%, 0.91% and 1.59% respectively (see bottom half of Table 1). We do not see improvements on the highly-challenging ANLI dataset, where multiple sentences were used for each premise.

To ensure that these improvements are not simply caused by regularization from supervising the attention weights, we create a randomised baseline by shuffling our desired distribution *D*, doing this separately for the premise and hypothesis. This highlights the effect of the supervision but without the additional information from the explanations. We find that this randomised baseline performs worse than the baseline with no supervision (89.50% accuracy on SNLI-test), with lower performance also seen on SNLI-hard (78.84%) and the MNLI datasets (71.5% and 71.23%).

When introducing an additional attention layer, the model with this extra layer does not outperform the baseline if the additional layer is not supervised. We therefore compare the supervised additional attention layer to our baseline without this additional layer. Supervising the additional attention layer significantly improves in-distribution performance with further improvements on SNLI-hard and MNLI (see the

	SNLI	Δ	MNLI	Δ	SNLI-hard	Δ	Params.
BERT Baseline	89.77		72.40		79.36		109m
LIREx-adapted	**90.79**	**+1.02**†	71.55	-0.85†	79.39	+0.03	453m
Pruthi et al-adapted.	89.99	+0.22†	73.27	+0.87†	79.90	+0.54†	109m
Ours (extra layer)	90.09	+0.35†	73.06	+0.67†	79.96	+0.60†	109m
Ours (existing attention)	90.17	+0.40†	**73.28**	**+0.88**†	**80.15**	**+0.79**†	109m

Table 2: Accuracy improvements compared to previous work, adapting Pruthi et al. (2020) for NLI and adapting LIREx (Zhao and Vydiswaran 2021) to use BERT models instead of the three RoBERTa models in its pipeline. † indicates statistically significant results compared to the baseline. Our methods and the Pruthi et al. (2020) method were tested over the same 25 random seeds, while the highly computationally expensive LIREx-adapted approach was evaluated over 5 random seeds.

top half of Table 1). While these results are also promising, we focus the remainder of the paper on supervising existing attention layers where we see greater improvements.

The in-distribution benefits from training with the explanations contrast with previous work on model robustness, with most work involving a trade-off between robustness and in-distribution performance (Sanh et al. 2020; Mahabadi, Belinkov, and Henderson 2020; Belinkov et al. 2019a). While some prior work retains in-distribution performance (Utama, Moosavi, and Gurevych 2020a), we find that training with explanations improves both in-distribution and out-of-distribution performance.

Experiments with DeBERTa

We evaluate the effect of training with explanations for DeBERTa, assessing whether the human explanations can improve even more powerful NLI models. We find that DeBERTa itself achieves 92.59% accuracy, outperforming previous state of the art results on SNLI (Zhang et al. 2020; Pilault, Elhattami, and Pal 2021; Sun et al. 2020). Combining the human explanations with DeBERTa provides a further statistically significant improvement for in-distribution performance, with the supervised model achieving 92.69% performance, a new state of the art result for SNLI. While the absolute improvement is small (0.1% for DeBERTa compared to 0.40% for BERT), it is more challenging to achieve as the potential room for improvement has also decreased.

Comparing Results with Prior Work

Our approach supervising existing model attention layers outperforms previously reported improvements, increasing SNLI performance by 0.40%. This compares to LIREx (Zhao and Vydiswaran 2021) which reported a 0.32% improvement in SNLI accuracy when training with a pipeline of three RoBERTa models and a GPT2 model. We recreate this result (LIREx-adapted), replacing the RoBERTa models in the pipeline with BERT models, then compare it to our BERT baseline (Table 2). As previous work using e-InferSent (Camburu et al. 2018), TextCat (Hase and Bansal 2021) and NILE (Kumar and Talukdar 2020) found no significant improvements using the explanations, we do not recreate these baselines. We find that LIREx-adapted has the largest improvement compared to the BERT baseline (+1.02%). This is unsurprising given that LIREx consists of a pipeline of four separate models, with a total of 453m parameters, compared to 109m parameters in the BERT baseline. In contrast, our approach of supervising an existing attention layer does not increase the number of parameters. LIREx-adapted also has a substantially lower performance than our DeBERTa model supervised with the explanations (90.79% for SNLI-test compared to 92.69%), despite using more parameters (453m compared to 409m).

Explanation type	Dev accuracy	Δ
Baseline	89.89	
Free text explanation	90.35	+0.46
Highlighted words	90.41	+0.52
Combined performance	**90.46**	+0.57

Table 3: Performance improvements were observed either when using free-text explanations or highlighted words, with the greatest improvements using a combination of these. Dev. accuracy is an average from 5 random seeds.

No previous work has shown out-of-distribution improvements from training with the explanations, and this continues to be the case with LIREx-adapted: the SNLI improvements for LIREx-adapted are accompanied by a fall in MNLI performance (-0.85), and almost no change in the SNLI-hard performance (Table 2).

We additionally show that adapting the approach presented by Pruthi et al. (2020) for NLI can also improve performance, with improvements across SNLI, MNLI and SNLI-hard. However, while improvements on MNLI are similar to our approach, improvements in SNLI-test are about half of the improvements we observed.

Choosing Which Explanations to Use and Which Heads to Supervise

We investigate different ways to use the e-SNLI explanations, assessing whether it is better to use the free-text explanations or the highlighted words. We also assess which attention heads should be supervised during training.

We find the best performance when combining both the free text explanations and the highlighted words within e-

	Premise			Hypothesis		
	P	R	F1	P	R	F1
Supervised LSTM-CRF (Thorne et al. 2019)	86.91	40.98	55.70	81.16	54.79	65.41
Unsupervised attention threshold (Thorne et al. 2019)	19.23	26.21	22.18	53.38	62.97	57.78
LIME (Thorne et al. 2019)	60.56	48.28	53.72	57.04	66.92	61.58
SE-NLI (Kim, Jang, and Allan 2020)	52.5	72.6	**60.9**	49.2	100.0	66.0
Baseline, with no supervision	0.51	0.01	0.03	43.32	58.65	49.83
Ours (existing attention)	55.20	58.60	56.85	61.48	78.96	**69.13**

Table 4: Precision, recall and F1 scores from token level predictions, using average attention values from 3 supervised attention heads. This is compared to a supervised LSTM-CRF model, LIME, SE-NLI, and the unsupervised attention approach.

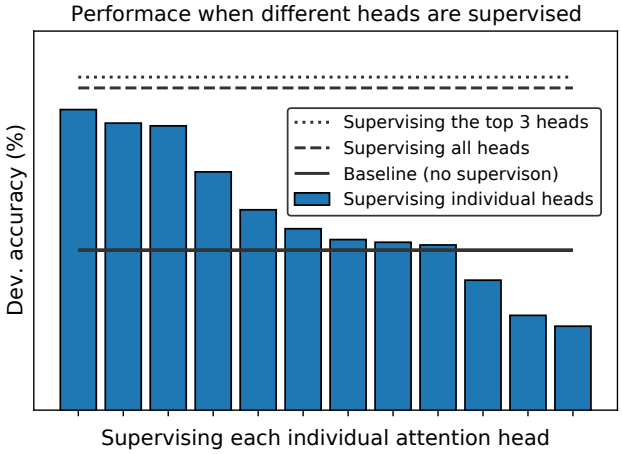

Figure 3: Accuracy when supervising each of the attention heads in turn, compared to the baseline with no supervision, supervising all heads and supervising the top 3 heads.

SNLI, taking an average of their attention distributions, $D_{freetext}$ and $D_{highlights}$ (see Table 3). When there are only words highlighted in the hypothesis for $D_{highlights}$, the attention is supervised using $D_{freetext}$, encouraging the model to pay attention to both sentences.

While we show that supervising all attention heads results in performance improvements (Figure 3), we find the best performance when only supervising 3 attention heads. This demonstrates how the additional supervision is only helpful for some attention heads, depending on the role of that specific head. Multi-head attention is designed to allow each head to perform a different function, therefore supervising all of them in the same direction can potentially have adverse effects. Figure 3 shows that the top 3 heads clearly performed better than the remaining heads when supervised individually, suggesting why this was the optimal number.

Analysis

Token Level Classification

To measure how successful the supervised heads are at identifying words in the human explanations, we consider the task of predicting which words appear in the highlighted explanations. The token-level classification is achieved by applying a threshold to the supervised attention weights, predicting whether a token is highlighted or not within e-SNLI. Unlike Thorne et al. (2019), Rei and Søgaard (2018) and Bujel, Yannakoudakis, and Rei (2021), we apply the token level thresholds to the normalised attention weights instead of the unnormalised weights, finding that this improves performance.

The model's token level predictions outperform a LSTM-CRF model jointly supervised for NLI and the token level task (Thorne et al. 2019; Lample et al. 2016) (see Table 4). We also compare this to an unsupervised approach using attention weights to make predictions (Thorne et al. 2019), LIME (Thorne et al. 2019; Ribeiro, Singh, and Guestrin 2016) and a perturbation-based self-explanation approach (Kim, Jang, and Allan 2020). The hypothesis F1 score for our approach is higher than previous baselines, with an improvement of 3.1 points. While Kim, Jang, and Allan (2020) find a higher F1 score for the premise, their work focused on improving the token level performance and did not improve the overall NLI task.

Understanding the Changes in Attention

To understand how the attention behaviour changes in our supervised model, we analyse the final [CLS] token attention compared to the baseline. The premise and the 1st [SEP] token only account for 22.86% of attention in the baseline, compared to 50.89% when supervising 12 heads. This highlights how the supervised model more evenly considers both the premise and hypothesis compared to the baseline.

Even in the earlier attention layers which were not directly supervised, more attention is paid to the premise in the supervised model (with 31.1% of attention in the baseline for the previous layer, compared to 54.2% with supervision). The increased focus on the premise may explain why performance is substantially better for SNLI-hard, a challenge set created from examples that a hypothesis-only model misclassified. Surprisingly, if we supervise only 3 heads in the top layer, lower layers attend to the premise to the same extent (with 54.8% of attention in the previous layer when supervising only 3 heads). This supports our decision to supervise fewer heads.

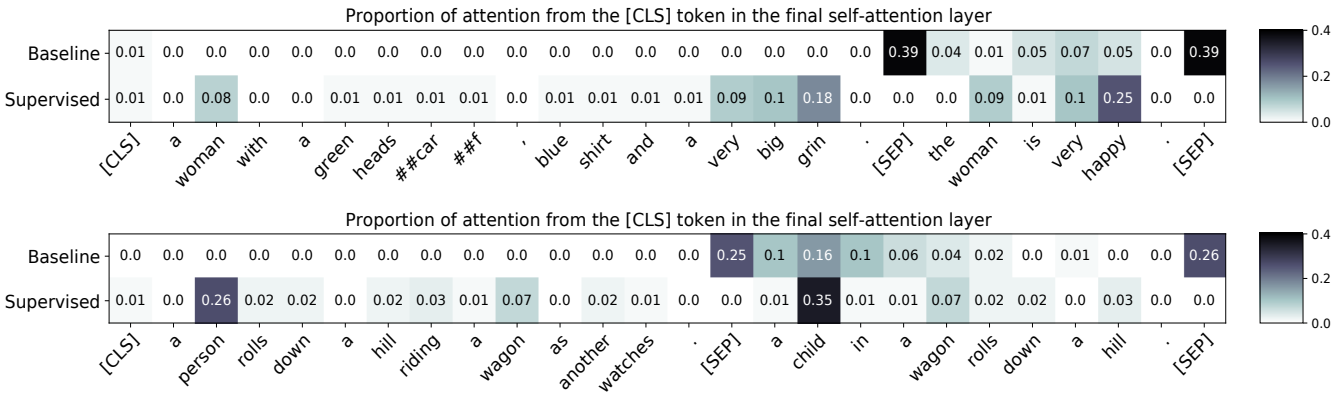

Figure 4: Average attention from the [CLS] token in the baseline and when we are supervising each attention head. Both models incorrectly predicted the first example as being neutral. The second example was correctly labeled by the supervised model (neutral), while the baseline model incorrectly predicted contradiction. The e-SNLI free-text explanations for the sentences include: 'One must be happy in order to have a big grin' and 'Just because it is a person does not mean it is a child'.

PoS Tag	12 heads	3 heads	Baseline
Noun	**54.3**	43.5	28.1
Verb	**20.4**	18.2	14.3
Adjective	**8.9**	8.3	5.2
Adposition	4.1	5.0	**7.8**
Determiner	3.4	6.0	**14.3**
Punctuation	0.9	7.7	**14.2**
Auxiliary	0.9	3.1	**8.2**
Other	7.1	8.2	7.9

Table 5: Percentage of attention across 5 seeds from the [CLS] token to tokens corresponding to different PoS tags.

Baseline		**Supervised**	
Words	%	Words	%
.	18.0	man	2.7
a	5.2	outside	2.5
is	4.0	woman	1.7
are	2.6	people	1.7
the	2.5	sitting	1.5

Table 6: Frequency in which each word is the most attended to token in a sentence pair across 5 random seeds.

Words Receiving Most Attention

In the supervised model, the words that receive the most attention are often nouns such as *man*, *woman*, or *people* (Table 6) which are the subjects of many sentences. Nouns are frequently used in the explanations, making up 46% of the highlighted words. On the other hand, stop-words are often attended to in the baseline, along with full-stops which may be a form of null attention (Vig and Belinkov 2019). More generally, using a SpaCy[3] Part of Speech tagger, after supervision we see less attention paid to punctuation, determiners and adposition words, while more attention is paid to nouns, verbs and adjectives (Table 5).

An analysis of the attention behaviour shows that the supervised model consistently attends to the most important words for the task, which is often not the case for the baseline model. In Figure 4, for each example the supervised model identifies the most important words in both the premise and the hypothesis. In the first sentence pair it attends to the word 'grin' in the premise and 'happy' in the hypothesis. In the second example, the supervised model identifies that the 'person' in the premise and 'child' in the hypothesis are the most important words.

Unlike the baseline, which mostly attends to the hypothesis and special tokens, the supervised model attends to words in the premise. As a result, the behaviour of the supervised model is more interpretable for NLI, where the class depends on the interaction between the two sentences.

Conclusion

Motivated by improving the robustness of NLI models based on human behaviour, we introduce a simple but effective approach that helps models learn from human explanations. We find the best performance when supervising a model's existing self-attention weights, encouraging more attention to be paid to words that are important in human explanations. Unlike prior work incorporating human explanations, our approach improves out-of-distribution performance alongside in-distribution performance, achieving a new state of the art result when combined with a DeBERTa model. Our supervised models have more interpretable attention weights and focus more on the most important words in each sentence, mostly nouns, verbs and adjectives. This contrasts with the baseline model that attends more to special tokens, stop-words and punctuation. The result is a model that attends to words humans believe are important, creating more robust and better performing NLI models.

[3]https://spacy.io

Acknowledgments

This research was partly supported by the ISRAEL SCIENCE FOUNDATION (grant No. 448/20). Y.B. was supported by an Azrieli Foundation Early Career Faculty Fellowship and by the Viterbi Fellowship in the Center for Computer Engineering at the Technion. We would like to thank the authors of the e-SNLI dataset for creating this excellent resource, and we also thank the LAMA reading group at Imperial for their feedback and encouragement.

References

Andreas, J.; Klein, D.; and Levine, S. 2018. Learning with Latent Language. In *NAACL*.

Belinkov, Y.; Poliak, A.; Shieber, S.; Van Durme, B.; and Rush, A. 2019a. Don't Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference. In *ACL*.

Belinkov, Y.; Poliak, A.; Shieber, S.; Van Durme, B.; and Rush, A. 2019b. On Adversarial Removal of Hypothesis-only Bias in Natural Language Inference. In *ACL*.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *EMNLP*.

Bujel, K.; Yannakoudakis, H.; and Rei, M. 2021. Zero-shot Sequence Labeling for Transformer-based Sentence Classifiers. In *RepL4NLP*.

Camburu, O.-M.; Rocktäschel, T.; Lukasiewicz, T.; and Blunsom, P. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *NeurIPS*.

Clark, C.; Yatskar, M.; and Zettlemoyer, L. 2019. Don't Take the Easy Way Out: Ensemble Based Methods for Avoiding Known Dataset Biases. In *EMNLP-IJCNLP*.

Clark, C.; Yatskar, M.; and Zettlemoyer, L. 2020. Learning to Model and Ignore Dataset Bias with Mixed Capacity Ensembles. In *EMNLP Findings*.

Clark, K.; Khandelwal, U.; Levy, O.; and Manning, C. D. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *BlackboxNLP@ACL*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. In *NAACL*.

Hase, P.; and Bansal, M. 2021. When Can Models Learn From Explanations? A Formal Framework for Understanding the Roles of Explanation Data. arXiv:2102.02201.

He, H.; Zha, S.; and Wang, H. 2019. Unlearn Dataset Bias in Natural Language Inference by Fitting the Residual. In *DeepLo@EMNLP*.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In *ICLR*.

Kim, Y.; Jang, M.; and Allan, J. 2020. Explaining Text Matching on Neural Natural Language Inference. *ACM Trans. Inf. Syst.*, 38(4).

Kumar, S.; and Talukdar, P. 2020. NILE : Natural Language Inference with Faithful Natural Language Explanations. In *ACL*. Online.

Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; and Dyer, C. 2016. Neural Architectures for Named Entity Recognition. In *NAACL*.

Liang, W.; Zou, J.; and Yu, Z. 2020. ALICE: Active Learning with Contrastive Natural Language Explanations. In *EMNLP*.

Liu, T.; Xin, Z.; Ding, X.; Chang, B.; and Sui, Z. 2020. An Empirical Study on Model-agnostic Debiasing Strategies for Robust Natural Language Inference. In *CoNLL*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mahabadi, R. K.; Belinkov, Y.; and Henderson, J. 2020. End-to-End Bias Mitigation by Modelling Biases in Corpora. In *ACL*.

Mahabadi, R. K.; Belinkov, Y.; and Henderson, J. 2021. Variational Information Bottleneck for Effective Low-Resource Fine-Tuning. In *ICLR*.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *AAAI*.

McCoy, T.; Pavlick, E.; and Linzen, T. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *ACL*.

Min, J.; McCoy, R. T.; Das, D.; Pitler, E.; and Linzen, T. 2020. Syntactic Data Augmentation Increases Robustness to Inference Heuristics. In *ACL*.

Minervini, P.; and Riedel, S. 2018. Adversarially Regularising Neural NLI Models to Integrate Logical Background Knowledge. In *CoNLL*.

Mu, J.; Liang, P.; and Goodman, N. 2020. Shaping Visual Representations with Language for Few-Shot Classification. In *ACL*.

Murty, S.; Koh, P. W.; and Liang, P. 2020. ExpBERT: Representation Engineering with Natural Language Explanations. In *ACL*.

Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; and Kiela, D. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *ACL*.

Pilault, J.; Elhattami, A.; and Pal, C. 2021. Conditionally Adaptive Multi-Task Learning: Improving Transfer Learning in NLP Using Fewer Parameters & Less Data. In *ICLR*.

Poliak, A.; Naradowsky, J.; Haldar, A.; Rudinger, R.; and Van Durme, B. 2018. Hypothesis Only Baselines in Natural Language Inference. In *SEM@NAACL*.

Pruthi, D.; Dhingra, B.; Soares, L. B.; Collins, M.; Lipton, Z. C.; Neubig, G.; and Cohen, W. W. 2020. Evaluating Explanations: How much do explanations from the teacher aid students? arXiv:2012.00893.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rajani, N. F.; McCann, B.; Xiong, C.; and Socher, R. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *ACL*.

Rei, M.; and Søgaard, A. 2018. Zero-Shot Sequence Labeling: Transferring Knowledge from Sentences to Tokens. In Walker, M. A.; Ji, H.; and Stent, A., eds., *NAACL*.

Rei, M.; and Søgaard, A. 2019. Jointly Learning to Label Sentences and Tokens. In *AAAI*.

Ribeiro, M.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *NAACL*.

Sanh, V.; Wolf, T.; Belinkov, Y.; and Rush, A. M. 2020. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *ICLR*.

Stacey, J.; Minervini, P.; Dubossarsky, H.; Riedel, S.; and Rocktäschel, T. 2020. Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training. In *EMNLP*.

Sun, Z.; Fan, C.; Han, Q.; Sun, X.; Meng, Y.; Wu, F.; and Li, J. 2020. Self-Explaining Structures Improve NLP Models. arXiv:2012.01786.

Teney, D.; Abbasnedjad, E.; and van den Hengel, A. 2020. Learning What Makes a Difference from Counterfactual Examples and Gradient Supervision. arXiv:2004.09034.

Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; and Mittal, A. 2019. Generating Token-Level Explanations for Natural Language Inference. In *NAACL*.

Tsuchiya, M. 2018. Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment. In *LREC*.

Tu, L.; Lalwani, G.; Gella, S.; and He, H. 2020. An Empirical Study on Robustness to Spurious Correlations using Pre-trained Language Models. *TACL*, 8: 621–633.

Utama, P. A.; Moosavi, N. S.; and Gurevych, I. 2020a. Mind the Trade-off: Debiasing NLU Models without Degrading the In-distribution Performance. In *ACL*.

Utama, P. A.; Moosavi, N. S.; and Gurevych, I. 2020b. Towards Debiasing NLU Models from Unknown Biases. In *EMNLP*. Online.

Vig, J.; and Belinkov, Y. 2019. Analyzing the Structure of Attention in a Transformer Language Model. In *BlackboxNLP@ACL*.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *NAACL*.

Yaghoobzadeh, Y.; Mehri, S.; Tachet des Combes, R.; Hazen, T. J.; and Sordoni, A. 2021. Increasing Robustness to Spurious Correlations using Forgettable Examples. In *EACL*.

Zhang, Z.; Wu, Y.; Zhao, H.; Li, Z.; Zhang, S.; Zhou, X.; and Zhou, X. 2020. Semantics-Aware BERT for Language Understanding. In *AAAI*.

Zhao, X.; and Vydiswaran, V. G. V. 2021. LIREx: Augmenting Language Inference with Relevant Explanation. In *AAAI*.