Attention is not Explanation

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Abstract

Attention mechanisms have seen wide adoption in neural NLP models. In addition to improving predictive performance, these are often touted as affording transparency: models equipped with attention provide a distribution over attended-to input units, and this is often presented (at least implicitly) as communicating the relative importance of inputs. However, it is unclear what relationship exists between attention weights and model outputs. In this work we perform extensive experiments across a variety of NLP tasks that aim to assess the degree to which attention weights provide meaningful "explanations" for predictions. We find that they largely do not. For example, learned attention weights are frequently uncorrelated with gradient-based measures of feature importance, and one can identify very different attention distributions that nonetheless yield equivalent predictions. Our findings show that standard attention modules do not provide meaningful explanations and should not be treated as though they do. Code to reproduce all experiments is available at https://github.com/successar/AttentionExplanation.

1 Introduction and Motivation

Attention mechanisms (Bahdanau et al., 2014) induce conditional distributions over input units to compose a weighted context vector for downstream modules. These are now a near-ubiquitous component of neural NLP architectures. Attention weights are often claimed (implicitly or explicitly) to afford insights into the "inner-workings" of models: for a given output one can inspect the inputs to which the model assigned large attention weights. Li *et al.* (2016) summarized this commonly held view in NLP: "Attention provides an important way to explain the workings of neural models". Indeed, claims that attention provides

after 15 minutes watching the movie i was *asking* myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a *waste* of time maybe i am not a 5 years old kid anymore

after 15 minutes watching the movie i was asking *myself* what to do leave the theater sleep or try to keep watching the movie to see if there *was* anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

original α
$f(x|\alpha, \theta) = 0.01$

adversarial $\tilde{\alpha}$
$f(x|\tilde{\alpha}, \theta) = 0.01$

Figure 1: Heatmap of attention weights induced over a negative movie review. We show observed model attention (left) and an adversarially constructed set of attention weights (right). Despite being quite dissimilar, these both yield effectively the same prediction (0.01).

interpretability are common in the literature, e.g., (Xu et al., 2015; Choi et al., 2016; Lei et al., 2017; Martins and Astudillo, 2016; Xie et al., 2017).[1]

Implicit in this is the assumption that the input units (e.g., words) accorded high attention weights are responsible for model outputs. But as far as we are aware, this assumption has not been formally evaluated, and our findings here suggest that it is problematic. More specifically, we empirically investigate the relationship between attention weights, inputs, and outputs. Assuming attention provides an explanation for model predictions, we might expect the following properties to hold. (i) Attention weights should correlate with feature importance measures (e.g., gradient-based measures); (ii) Alternative (or *counterfactual*) attention weight configurations ought to yield corresponding changes in prediction (and if they do not then are equally plausible as explanations). We report that neither property is consistently observed by standard attention mechanisms in the context of text classification, question answering (QA), and Natural Language Inference (NLI) tasks when RNN encoders are used.

---

[1]We do not intend to single out any particular work; indeed one of the authors has himself presented (supervised) attention as providing interpretability (Zhang et al., 2016).

Consider Figure 1. The left panel shows the original attention distribution α over the words of a particular movie review using a standard attentive BiLSTM architecture for sentiment analysis. It is tempting to conclude from this that the token *waste* is largely responsible for the model coming to its disposition of 'negative' ($\hat{y} = 0.01$). But one can construct an alternative attention distribution $\tilde{\alpha}$ (right panel) that attends to entirely different tokens yet yields an essentially identical prediction (holding all other parameters of f, θ, constant).

Such counterfactual distributions imply that explaining the original prediction by highlighting attended-to tokens is misleading insofar as alternative attention distributions would have yielded an equivalent prediction (e.g., one might conclude from the right panel that model output was due primarily to *was* rather than *waste*). Further, the attention weights in this case correlate only weakly with gradient-based measures of feature importance ($\tau_g = 0.29$). And arbitrarily permuting the entries in α yields a median output difference of 0.006 with the original prediction.

These and similar findings call into question the view that attention provides meaningful insight into model predictions. We thus caution against using attention weights to highlight input tokens "responsible for" model outputs and constructing just-so stories on this basis, particularly with complex encoders.

Research questions and contributions. We examine the extent to which the (often implicit) narrative that attention provides model *transparency*[2] holds across tasks by exploring the following empirical questions.

1. To what extent do induced attention weights correlate with measures of feature importance – specifically, those resulting from gradients and *leave-one-out* methods?

2. Would alternative attention weights (and hence distinct heatmaps/"explanations") necessarily yield different predictions?

Our findings with respect to these questions (assuming a BiRNN encoder) are summarized as follows: (1) Only weakly and inconsistently, and, (2) No; it is very often possible to construct *adversarial* attention distributions that yield effectively

equivalent predictions as when using the originally induced attention weights, despite attending to entirely different input features. Further, randomly permuting attention weights often induces only minimal changes in output.

2 Preliminaries and Assumptions

We consider exemplar NLP tasks for which attention mechanisms are commonly used: classification, natural language inference (NLI), and question answering.[3] We adopt the following general modeling assumptions and notation.

We assume model inputs $\mathbf{x} \in \mathbb{R}^{T \times |V|}$, composed of one-hot encoded words at each position. These are passed through an embedding matrix $\mathbf{E}$ which provides dense (d dimensional) token representations $\mathbf{x}_e \in \mathbb{R}^{T \times d}$. Next, an encoder **Enc** consumes the embedded tokens in order, producing T m-dimensional hidden states: $\mathbf{h} = \mathbf{Enc}(\mathbf{x}_e) \in \mathbb{R}^{T \times m}$. We predominantly consider a Bi-RNN as the encoder module, but for completeness we also analyze convolutional and (unordered) 'average embedding' variants.[4]

A similarity function ϕ maps $\mathbf{h}$ and a query $\mathbf{Q} \in \mathbb{R}^m$ (e.g., hidden representation of a question in QA, or the hypothesis in NLI) to scalar scores, and attention is then induced over these: $\hat{\alpha} = \text{softmax}(\phi(\mathbf{h}, \mathbf{Q})) \in \mathbb{R}^T$. In this work we consider two common similarity functions: *Additive* $\phi(\mathbf{h}, \mathbf{Q}) = \mathbf{v}^T \tanh(\mathbf{W_1 h} + \mathbf{W_2 Q})$ (Bahdanau et al., 2014) and *Scaled Dot-Product* $\phi(\mathbf{h}, \mathbf{Q}) = \frac{\mathbf{hQ}}{\sqrt{m}}$ (Vaswani et al., 2017), where $\mathbf{v}, \mathbf{W_1}, \mathbf{W_2}$ are model parameters.

Finally, a dense layer **Dec** with parameters $\boldsymbol{\theta}$ consumes a weighted instance representation and yields a prediction $\hat{y} = \sigma(\boldsymbol{\theta} \cdot h_\alpha) \in \mathbb{R}^{|\mathcal{Y}|}$, where $h_\alpha = \sum_{t=1}^{T} \hat{\alpha}_t \cdot h_t$; σ is an output activation function; and $|\mathcal{Y}|$ denotes the label set size.

3 Datasets and Tasks

For **binary text classification**, we use:

Stanford Sentiment Treebank (SST) (Socher et al., 2013). 10,662 sentences tagged with sentiment on a scale from 1 (most negative) to 5 (most positive). We filter out neutral instances and dichotomize the remaining sentences into positive (4, 5) and negative (1, 2).

---

[2]Defined as per (Lipton, 2016); we are interested in whether attended-to features are responsible for outputs.

[3]While attention is perhaps *most* common in seq2seq tasks like translation, our impression is that interpretability is not typically emphasized for such tasks, in general.

[4]In the latter case, h_t is the embedding of token t after being passed through a linear layer and ReLU activation.

Dataset	$|V|$	Avg. length	Train size	Test size	Test performance
SST	16175	19	3034 / 3321	863 / 862	0.81
IMDB	13916	179	12500 / 12500	2184 / 2172	0.88
ADR Tweets	8686	20	14446 / 1939	3636 / 487	0.61
20 Newsgroups	8853	115	716 / 710	151 / 183	0.94
AG News	14752	36	30000 / 30000	1900 / 1900	0.96
Diabetes (MIMIC)	22316	1858	6381 / 1353	1295 / 319	0.79
Anemia (MIMIC)	19743	2188	1847 / 3251	460 / 802	0.92
CNN	74790	761	380298	3198	0.64
bAbI (Task 1 / 2 / 3)	40	8 / 67 / 421	10000	1000	1.0 / 0.65 / 0.64
SNLI	20982	14	182764 / 183187 / 183416	3219 / 3237 / 3368	0.78

Table 1: Dataset characteristics. For train and test size, we list the cardinality for each class, where applicable: 0/1 for binary classification (top), and 0 / 1 / 2 for NLI (bottom). Average length is in tokens. Test metrics are F1 score, accuracy, and micro-F1 for classification, QA, and NLI, respectively; all correspond to performance using a BiLSTM encoder. We note that results using convolutional and average (i.e., non-recurrent) encoders are comparable for classification though markedly worse for QA tasks.

IMDB Large Movie Reviews Corpus (Maas et al., 2011). Binary sentiment classification dataset containing 50,000 polarized (positive or negative) movie reviews, split into half for training and testing.

Twitter Adverse Drug Reaction dataset (Nikfarjam et al., 2015). A corpus of ∼8000 tweets retrieved from Twitter, annotated by domain experts as mentioning adverse drug reactions.

20 Newsgroups (Hockey vs Baseball). Collection of ∼20,000 newsgroup correspondences, partitioned (nearly) evenly across 20 categories. We extract instances belonging to *baseball* and *hockey*, which we designate as 0 and 1, respectively, to derive a binary classification task.

AG News Corpus (Business vs World).[5] 496,835 news articles from 2000+ sources. We follow (Zhang et al., 2015) in filtering out all but the top 4 categories. We consider the binary classification task of discriminating between *world* (0) and *business* (1) articles.

MIMIC ICD9 (Diabetes) (Johnson et al., 2016). A subset of discharge summaries from the MIMIC III dataset of electronic health records. The task is to recognize if a given summary has been labeled with the ICD9 code for diabetes (or not).

MIMIC ICD9 (Chronic vs Acute Anemia) (Johnson et al., 2016). A subset of discharge summaries from MIMIC III dataset (Johnson et al., 2016) known to correspond to patients with anemia. Here the task to distinguish the type of anemia for each report – *acute* (0) or *chronic* (1).

For **Question Answering (QA)**:

CNN News Articles (Hermann et al., 2015). A corpus of cloze-style questions created via auto-

matic parsing of news articles from CNN. Each instance comprises a paragraph-question-answer triplet, where the answer is one of the anonymized entities in the paragraph.

bAbI (Weston et al., 2015). We consider the three tasks presented in the original bAbI dataset paper, training separate models for each. These entail finding (i) a single supporting fact for a question and (ii) two or (iii) three supporting statements, chained together to compose a coherent line of reasoning.

Finally, for **Natural Language Inference** (NLI):

The *SNLI dataset* (Bowman et al., 2015). 570k human-written English sentence pairs manually labeled for balanced classification with the labels *neutral*, *contradiction*, and *entailment*, supporting the task of natural language inference (NLI). In this work, we generate an attention distribution over premise words conditioned on the hidden representation induced for the hypothesis.

We restrict ourselves to comparatively simple instantiations of attention mechanisms, as described in the preceding section. This means we do not consider recently proposed 'BiAttentive' architectures that attend to tokens in the respective inputs, conditioned on the other inputs (Parikh et al., 2016; Seo et al., 2016; Xiong et al., 2016).

Table 1 provides summary statistics for all datasets, as well as the observed test performances for additional context.

4 Experiments

We run a battery of experiments that aim to examine empirical properties of learned attention weights and to interrogate their interpretability and transparency. The key questions are: *Do*

---

[5] http://www.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

Dataset	Class	Gradient (BiLSTM) τ_g		Gradient (Average) τ_g		Leave-One-Out (BiLSTM) τ_{loo}	
		Mean ± Std.	Sig. Frac.	Mean ± Std.	Sig. Frac.	Mean ± Std.	Sig. Frac.
SST	0	0.34 ± 0.21	0.48	0.61 ± 0.20	0.87	0.27 ± 0.19	0.33
	1	0.36 ± 0.21	0.49	0.60 ± 0.21	0.83	0.32 ± 0.19	0.40
IMDB	0	0.44 ± 0.06	1.00	0.67 ± 0.05	1.00	0.34 ± 0.07	1.00
	1	0.43 ± 0.06	1.00	0.68 ± 0.05	1.00	0.34 ± 0.07	0.99
ADR Tweets	0	0.47 ± 0.18	0.76	0.73 ± 0.13	0.96	0.29 ± 0.20	0.44
	1	0.49 ± 0.15	0.85	0.72 ± 0.12	0.97	0.44 ± 0.16	0.74
20News	0	0.07 ± 0.17	0.37	0.79 ± 0.07	1.00	0.06 ± 0.15	0.29
	1	0.21 ± 0.22	0.61	0.75 ± 0.08	1.00	0.20 ± 0.20	0.62
AG News	0	0.36 ± 0.13	0.82	0.78 ± 0.07	1.00	0.30 ± 0.13	0.69
	1	0.42 ± 0.13	0.90	0.76 ± 0.07	1.00	0.43 ± 0.14	0.91
Diabetes	0	0.42 ± 0.05	1.00	0.75 ± 0.02	1.00	0.41 ± 0.05	1.00
	1	0.40 ± 0.05	1.00	0.75 ± 0.02	1.00	0.45 ± 0.05	1.00
Anemia	0	0.47 ± 0.05	1.00	0.77 ± 0.02	1.00	0.46 ± 0.05	1.00
	1	0.46 ± 0.06	1.00	0.77 ± 0.03	1.00	0.47 ± 0.06	1.00
CNN	Overall	0.24 ± 0.07	0.99	0.50 ± 0.10	1.00	0.20 ± 0.07	0.98
bAbI 1	Overall	0.25 ± 0.16	0.55	0.72 ± 0.12	0.99	0.16 ± 0.14	0.28
bAbI 2	Overall	−0.02 ± 0.14	0.27	0.68 ± 0.06	1.00	−0.01 ± 0.13	0.27
bAbI 3	Overall	0.24 ± 0.11	0.87	0.61 ± 0.13	1.00	0.26 ± 0.10	0.89
SNLI	0	0.31 ± 0.23	0.36	0.59 ± 0.18	0.80	0.16 ± 0.26	0.20
	1	0.33 ± 0.21	0.38	0.58 ± 0.19	0.80	0.36 ± 0.19	0.44
	2	0.31 ± 0.21	0.36	0.57 ± 0.19	0.80	0.34 ± 0.20	0.40

Table 2: Mean and std. dev. of correlations between gradient/leave-one-out importance measures and attention weights. *Sig. Frac.* columns report the fraction of instances for which this correlation is statistically significant; note that this largely depends on input length, as correlation does tend to exist, just weakly. Encoders are denoted parenthetically. These are representative results; exhaustive results for all encoders are available to browse online.

learned attention weights agree with alternative, natural measures of feature importance? And, *Had we attended to different features, would the prediction have been different*?

More specifically, in Section 4.1, we empirically analyze the correlation between gradient-based feature importance and learned attention weights, and between 'leave-one-out' (LOO) measures and the same. In Section 4.2 we then consider counterfactual (to those observed) attention distributions. Under the assumption that attention weights are explanatory, such counterfactual distributions may be viewed as alternative potential explanations; if these do not correspondingly change model output, then the original attention weights do not provide unique explanation for predictions, i.e., attending to other features could have resulted in the same output.

To generate counterfactual attention distributions, we first consider randomly permuting observed attention weights and recording associated changes in model outputs (4.2.1). We then propose explicitly searching for "adversarial" attention weights that maximally differ from the observed attention weights (which one might show in a heatmap and use to explain a model prediction), and yet yield an effectively equivalent prediction (4.2.2). The latter strategy also provides a useful potential metric for the reliability of attention weights as explanations: we can report a measure

quantifying how different attention weights can be for a given instance without changing the model output by more than some threshold ϵ.

All results presented below are generated on test sets. We present results for *Additive* attention below. The results for *Scaled Dot Product* in its place are comparable. We provide a web interface to interactively browse the (very large set of) plots for all datasets, model variants, and experiment types: https://successar.github.io/AttentionExplanation/docs/.

In the following sections, we use Total Variation Distance (TVD) as the measure of change between output distributions, defined as follows. $\text{TVD}(\hat{y}_1, \hat{y}_2) = \frac{1}{2} \sum_{i=1}^{|\mathcal{Y}|} |\hat{y}_{1i} - \hat{y}_{2i}|$. We use the Jensen-Shannon Divergence (JSD) to quantify the difference between two attention distributions: $\text{JSD}(\alpha_1, \alpha_2) = \frac{1}{2}\text{KL}[\alpha_1 || \frac{\alpha_1 + \alpha_2}{2}] + \frac{1}{2}\text{KL}[\alpha_2 || \frac{\alpha_1 + \alpha_2}{2}]$.

4.1 Correlation Between Attention and Feature Importance Measures

We empirically characterize the relationship between attention weights and corresponding feature importance scores. Specifically we measure correlations between attention and: (1) gradient based measures of feature importance (τ_g), and, (2) differences in model output induced by leaving features out (τ_{loo}). While these measures are themselves insufficient for interpretation of neu-

(a) SST (BiLSTM)   (b) SST (Average)   (c) Diabetes (BiLSTM)   (d) Diabetes (Average)

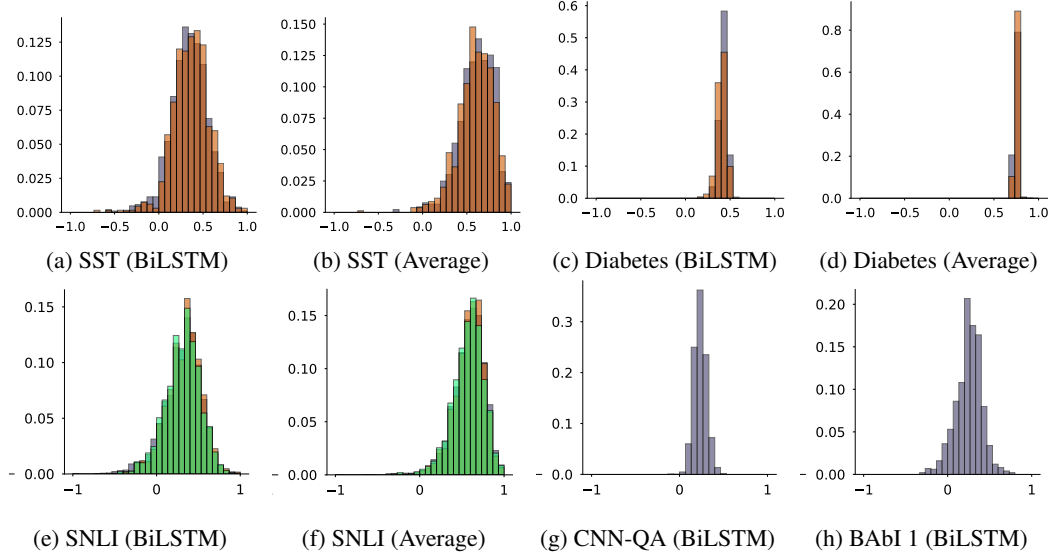(e) SNLI (BiLSTM)   (f) SNLI (Average)   (g) CNN-QA (BiLSTM)   (h) BAbI 1 (BiLSTM)

Figure 2: Histogram of Kendall τ between attention and gradients. Encoder variants are denoted parenthetically; colors indicate predicted classes. Exhaustive results are available for perusal online. Best viewed in color.

ral model behavior (Feng et al., 2018), they do provide measures of individual feature importance with known semantics (Ross et al., 2017). It is thus instructive to ask whether these measures correlate with attention weights.

The process we follow to quantify this is described in Algorithm 1. We denote the input resulting from removing the word at position t in $\mathbf{x}$ by $\mathbf{x}_{-t}$. Note that we disconnect the computation graph at the attention module so that the gradient does not flow through this layer.

Algorithm 1 Feature Importance Computations

$\mathbf{h} \leftarrow \text{Enc}(\mathbf{x}), \hat{\alpha} \leftarrow \text{softmax}(\phi(\mathbf{h}, \mathbf{Q}))$
$\hat{y} \leftarrow \text{Dec}(\mathbf{h}, \alpha)$
$g_t \leftarrow |\sum_{w=1}^{|V|} \mathbb{1}[\mathbf{x}_{tw} = 1] \frac{\partial y}{\partial \mathbf{x}_{tw}}|, \forall t \in [1, T]$
$\tau_g \leftarrow \text{Kendall-}\tau(\alpha, g)$
$\Delta \hat{y}_t \leftarrow \text{TVD}(\hat{y}(\mathbf{x}_{-t}), \hat{y}(\mathbf{x})), \forall t \in [1, T]$
$\tau_{loo} \leftarrow \text{Kendall-}\tau(\alpha, \Delta \hat{y})$

Table 2 reports summary statistics of Kendall τ correlations for each dataset. Full distributions are shown in Figure 2, which plots histograms of τ_g for every data point in the respective corpora. (Corresponding plots for τ_{loo} are similar and the full set can be browsed via the online supplement.) We plot these separately for each class: orange (■) represents instances predicted as positive, and purple (■) those predicted to be negative. For SNLI, colors ■, ■ and ■ code for contradiction, entailment, and neutral respectively.

In general, observed correlations are modest

(recall: 0 indicates no correspondence, 1 implies perfect concordance) for the BiRNN encoder. The centrality of observed densities hovers around or below 0.5 in most of the corpora considered. Moreover, as per Table 2, correlation is sufficiently weak that a statistically significant correlation between attention weights and feature importance scores (both gradient and feature erasure based) cannot consistently be established across corpora.

In contrast, gradients in "average" embedding based models show very high degree of correspondence with attention weights – on average across corpora, correlation between LOO scores and attention weights is ~0.375 points higher for this encoder, compared to the BiLSTM. These results suggest that, in general, attention weights do not strongly or consistently agree with such feature importance scores in models with contextualized embeddings. This is problematic for the view of attention weights as explanatory, given the face validity of input gradient/erasure based explanations (Ross et al., 2017; Li et al., 2016). On some datasets — notably the MIMIC tasks, and to a lesser extent the QA corpora — this correlation is consistently significant but remains relatively weak. This could be attributed to increased length of documents for these datasets providing stronger signal to standard hypothesis testing methods.

For reference we report correlations between gradients and LOO scores in the Appendix and online materials; these are consistently stronger than the correlation between attention weights and

3547

either feature importance score for the recurrent (BiLSTM) encoder. These exhibit, on average, a (i) 0.2 and (ii) ~ 0.25 greater correlation with each other than BiLSTM attention and (i) LOO and (ii) gradient scores.

4.2 Counterfactual Attention Weights

We next consider *what-if* scenarios corresponding to alternative (counterfactual) attention weights. The idea is to investigate whether the prediction would have been different, had the model emphasized (attended to) different input features. More precisely, suppose $\hat{\alpha} = \{\hat{\alpha}_t\}_{t=1}^T$ are the attention weights induced for an instance, giving rise to model output $\hat{y}$. We then consider counterfactual distributions over y, under alternative α.

We experiment with two means of constructing such distributions. First, we simply scramble the original attention weights $\hat{\alpha}$, re-assigning each value to an arbitrary, randomly sampled index (input feature). Second, we generate an *adversarial attention distribution*: this is a set of attention weights that is maximally distinct from $\hat{\alpha}$ but that nonetheless yields an equivalent prediction (i.e., prediction within some ϵ of $\hat{y}$).

4.2.1 Attention Permutation

To characterize model behavior when attention weights are shuffled, we follow Algorithm 2.

Algorithm 2 Permuting attention weights

$\quad \mathbf{h} \leftarrow \mathrm{Enc}(\mathbf{x}), \hat{\alpha} \leftarrow \mathrm{softmax}(\phi(\mathbf{h}, \mathbf{Q}))$
$\quad \hat{y} \leftarrow \mathrm{Dec}(\mathbf{h}, \hat{\alpha})$
$\quad \textbf{for } p \leftarrow 1 \text{ to } 100 \textbf{ do}$
$\qquad \alpha^p \leftarrow \mathrm{Permute}(\hat{\alpha})$
$\qquad \hat{y}^p \leftarrow \mathrm{Dec}(\mathbf{h}, \alpha^p) \qquad \triangleright \text{Note}: \mathbf{h} \text{ is not changed}$
$\qquad \Delta \hat{y}^p \leftarrow \mathrm{TVD}[\hat{y}^p, \hat{y}]$
$\quad \textbf{end for}$
$\quad \Delta \hat{y}^{med} \leftarrow \mathrm{Median}_p(\Delta \hat{y}^p)$

Figure 3 depicts the relationship between the maximum attention value in the original $\hat{\alpha}$ and the median induced change in model output ($\Delta \hat{y}^{med}$) across instances in the respective datasets. Colors again indicate class predictions, as above.

We observe that there exist many points with small $\Delta \hat{y}^{med}$ despite large magnitude attention weights. These are cases in which the attention weights might suggest explaining an output by a small set of features (this is how one might reasonably read a heatmap depicting the attention weights), but where scrambling the attention

makes little difference to the prediction.

In some cases, such as predicting ICD codes from notes using the MIMIC dataset, one can see different behavior for the respective classes. For the *Diabetes* task, e.g., attention behaves intuitively for at least the positive class; perturbing attention in this case causes large changes to the prediction. We again conjecture that this is due to a few tokens serving as high precision indicators for the positive class; in their absence (or when they are not attended to sufficiently), the prediction drops considerably. However, this is the exception rather than the rule.

4.2.2 Adversarial Attention

We next propose a more focused approach to counterfactual attention weights, which we will refer to as *adversarial attention*. The intuition is to explicitly seek out attention weights that differ as much as possible from the observed attention distribution and yet leave the prediction effectively unchanged. Such adversarial weights violate an intuitive property of explanations: shifting model attention to very different input features should yield corresponding changes in the output. Alternative attention distributions identified adversarially may then be viewed as equally plausible explanations for the same output.

Operationally, realizing this objective requires specifying a value ϵ that defines what qualifies as a "small" difference in model output. Once this is specified, we aim to find k adversarial distributions $\{\alpha^{(1)}, ..., \alpha^{(k)}\}$, such that each $\alpha^{(i)}$ maximizes the distance from original $\hat{\alpha}$ but does not change the output by more than ϵ. In practice we simply set this to 0.01 for text classification and 0.05 for QA datasets.[6]

We propose the following optimization problem to identify adversarial attention weights.

$$\underset{\alpha^{(1)}, ..., \alpha^{(k)}}{\mathrm{maximize}} \quad f(\{\alpha^{(i)}\}_{i=1}^k) \tag{1}$$
$$\text{subject to} \quad \forall i \; \mathrm{TVD}[\hat{y}(\mathbf{x}, \alpha^{(i)}), \hat{y}(\mathbf{x}, \hat{\alpha})] \leq \epsilon$$

Where $f(\{\alpha^{(i)}\}_{i=1}^k)$ is:

$$\sum_{i=1}^k \mathrm{JSD}[\alpha^{(i)}, \hat{\alpha}] + \frac{1}{k(k-1)} \sum_{i<j} \mathrm{JSD}[\alpha^{(i)}, \alpha^{(j)}] \tag{2}$$

---

[6]We make the threshold slightly higher for QA because the output space is larger and thus small dimension-wise perturbations can produce comparatively large TVD.
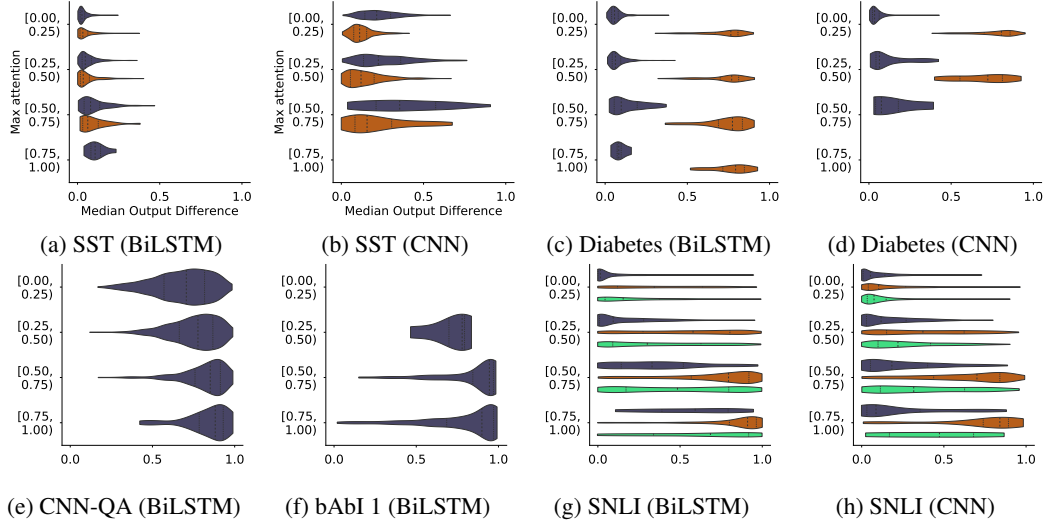
Figure 3: Median change in output ($\Delta \hat{y}^{med}$) (x-axis) densities in relation to the max attention ($\max \hat{\alpha}$) (y-axis) obtained by randomly permuting instance attention weights. Encoders denoted parenthetically. Plots for all corpora and using all encoders are available online.

In practice we maximize a relaxed version of this objective via the Adam SGD optimizer (Kingma and Ba, 2014): $f(\{\alpha^{(i)}\}_{i=1}^{k}) + \frac{\lambda}{k} \sum_{i=1}^{k} \max(0, \text{TVD}[\hat{y}(\mathbf{x}, \alpha^{(i)}), \hat{y}(\mathbf{x}, \hat{\alpha})] - \epsilon).[7]

Equation 1 attempts to identify a set of new attention distributions over the input that is as far as possible from the observed α (as measured by JSD) and from each other (and thus diverse), while keeping the output of the model within ϵ of the original prediction. We denote the output obtained under the ith adversarial attention by $\hat{y}^{(i)}$. Note that the JS Divergence between any two categorical distributions (irrespective of length) is bounded from above by 0.69.

One can view an attentive decoder as a function that maps from the space of latent input representations and attention weights over input words Δ^{T-1} to a distribution over the output space $\mathcal{Y}$. Thus, for any output $\hat{y}$, we can define how likely each attention distribution α will generate the output as inversely proportional to $\text{TVD}(y(\alpha), \hat{y})$.

Figure 4 depicts the distributions of max JSDs realized over instances with adversarial attention weights for a subset of the datasets considered. Colors again indicate predicted class. Mass toward the upper-bound of 0.69 indicates that we are frequently able to identify maximally different attention weights that hardly budge model output. We observe that one can identify adversarial attention weights associated with high JSD for a significant number of examples. This means that is often the

---

Algorithm 3 Finding adversarial attention weights

$\mathbf{h} \leftarrow \text{Enc}(\mathbf{x}), \hat{\alpha} \leftarrow \text{softmax}(\phi(\mathbf{h}, \mathbf{Q}))$
$\hat{y} \leftarrow \text{Dec}(\mathbf{h}, \hat{\alpha})$
$\alpha^{(1)}, ..., \alpha^{(k)} \leftarrow \text{Optimize Eq } 1$
for $i \leftarrow 1$ to k **do**
    $\hat{y}^{(i)} \leftarrow \text{Dec}(\mathbf{h}, \alpha^{(i)})$ ▷ $\mathbf{h}$ is not changed
    $\Delta \hat{y}^{(i)} \leftarrow \text{TVD}[\hat{y}, \hat{y}^{(i)}]$
    $\Delta \alpha^{(i)} \leftarrow \text{JSD}[\hat{\alpha}, \alpha^{(i)}]$
end for
$\epsilon\text{-max JSD} \leftarrow \max_i \mathbb{1}[\Delta \hat{y}^{(i)} \leq \epsilon]\Delta \alpha^{(i)}$

---

case that quite different attention distributions over inputs would yield essentially the same (within ϵ) output.

In the case of the *diabetes* task, we again observe a pattern of low JSD for positive examples (where evidence is present) and high JSD for negative examples. In other words, for this task, if one perturbs the attention weights when it is inferred that the patient is diabetic, this does change the output, which is intuitively agreeable. However, this behavior again is an exception to the rule.

We also consider the relationship between max attention weights (indicating strong emphasis on a particular feature) and the dissimilarity of identified adversarial attention weights, as measured via JSD, for adversaries that yield a prediction within ϵ of the original model output. Intuitively, one might hope that if attention weights are peaky, then counterfactual attention weights that are very different but which yield equivalent predictions
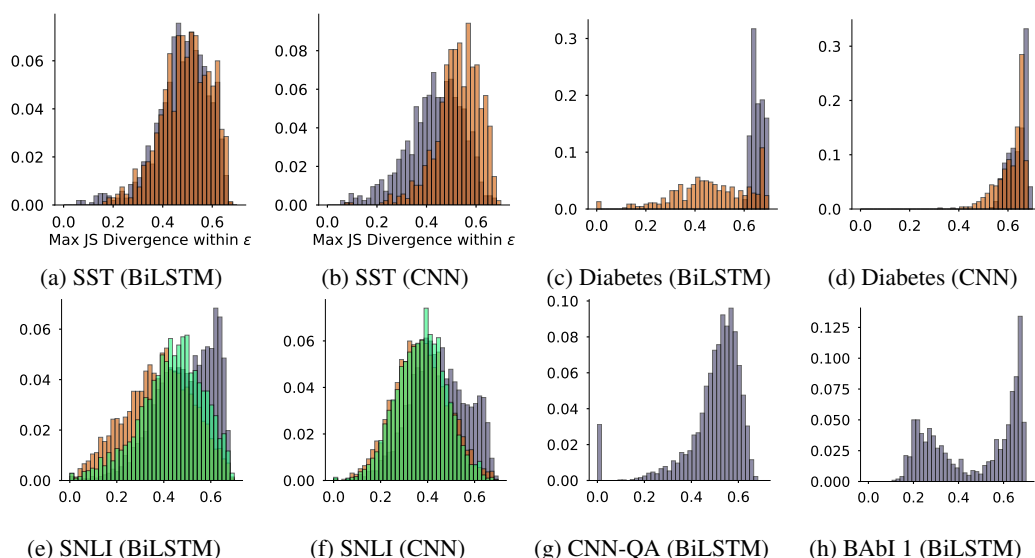
---

[7]We set $\lambda = 500$.

Figure 4: Histogram of maximum adversarial JS Divergence (ϵ-max JSD) between original and adversarial attentions over all instances. In all cases shown, $|\hat{y}^{adv} - \hat{y}| < \epsilon$. Encoders are specified in parantheses. Best viewed in color.

would be more difficult to identify.

Figure 5 illustrates that while there is a negative trend to this effect, it is realized only weakly. Put another way: there exist many cases (in all datasets) in which despite a high attention weight, an alternative and quite different attention configuration over inputs yields effectively the same output. In light of this, presenting a heatmap implying that a particular set of features is primarily responsible for an output would seem to be misleading.

5 Related Work

We have focused on attention mechanisms and the question of whether they afford transparency, but a number of interesting strategies unrelated to attention mechanisms have been recently proposed to provide insights into neural NLP models. These include approaches that measure feature importance based on gradient information (Ross et al., 2017; Sundararajan et al., 2017) (aligned with the gradient-based measures that we have used here), and methods based on *representation erasure* (Li et al., 2016), in which dimensions are removed and then the resultant change in output is recorded (similar to our experiments with removing tokens from inputs, albeit we do this at the input layer).

Comparing such importance measures to attention scores may provide additional insights into the working of attention based models (Ghaeini et al., 2018). Another novel line of work in this direction involves explicitly identifying explanations of black-box predictions via a causal frame-

work (Alvarez-Melis and Jaakkola, 2017). We also note that there has been complementary work demonstrating correlation between human attention and induced attention weights, which was relatively strong when humans agreed on an explanation (Pappas and Popescu-Belis, 2016). It would be interesting to explore if such cases present explicit 'high precision' signals in the text (for example, the positive label in diabetes dataset).

More specific to attention mechanisms, recent promising work has proposed more principled attention variants designed explicitly for interpretability; these may provide greater transparency by imposing *hard*, *sparse* attention. Such instantiations explicitly select (modest) subsets of inputs to be considered when making a prediction, which are then by construction responsible for model output (Lei et al., 2016; Peters et al., 2018). *Structured attention* models (Kim et al., 2017) provide a generalized framework for describing and fitting attention variants with explicit probabilistic semantics. Tying attention weights to human-provided rationales is another potentially promising avenue (Bao et al., 2018). We hope our work motivates further development of these methods, resulting in attention variants that both improve predictive performance and provide insights into model predictions.

6 Discussion and Conclusions

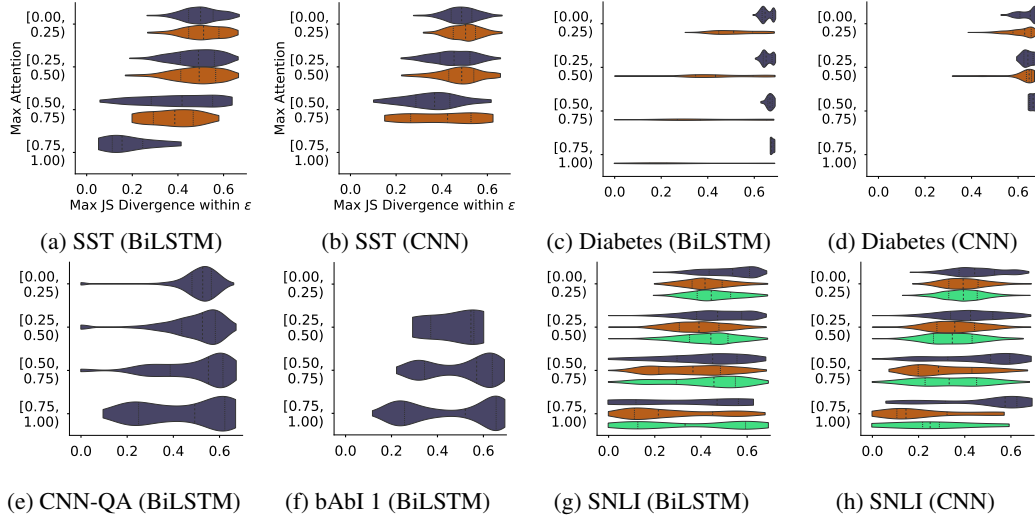We have provided evidence that correlation between intuitive feature importance measures (in-

Figure 5: Densities of maximum JS divergences (ϵ-max JSD) (x-axis) as a function of the max attention (y-axis) in each instance for obtained between original and adversarial attention weights.

cluding gradient and feature erasure approaches) and learned attention weights is weak when using a BiRNN encoder (Section 4.1). We also established that counterfactual attention distributions — which would tell a different story about why a model made the prediction that it did — often have no effect on model output (Section 4.2).

These results suggest that while attention modules consistently yield improved performance on NLP tasks, their ability to provide transparency for model predictions is (in the sense of pointing to inputs responsible for outputs) questionable. More generally, how one is meant to interpret the 'heatmaps' of attention weights placed over inputs that are commonly presented is unclear. These seem to suggest a story about how a model arrived at a particular disposition, but the results here indicate that the relationship between this and attention is not obvious, at least for RNN encoders.

There are important **limitations** to this work and the conclusions we can draw from it. We have reported the (generally weak) correlation between learned attention weights and various alternative measures of feature importance, e.g., gradients. We do not imply that such alternative measures are necessarily ideal or should be considered 'ground truth'. While such measures do enjoy a clear intrinsic (to the model) semantics, their interpretation for non-linear neural networks can nonetheless be difficult for humans (Feng et al., 2018). Still, that attention consistently correlates poorly with *multiple* such measures ought to give pause to practitioners. That said, exactly how strong such correlations 'should' be to establish reliability as

explanation is an admittedly subjective question.

We note that the counterfactual attention experiments demonstrate the existence of alternative heatmaps that yield equivalent predictions; thus one cannot conclude that the model made a particular prediction *because* it attended over inputs in a specific way. But these adversarial weights may themselves be unlikely under the attention module parameters. Further, it may be that multiple plausible explanations exist, complicating interpretation. We would maintain that in such cases the model should highlight all plausible explanations, but one may instead view a model that provides 'sufficient' explanation as reasonable.

An additional limitation is that we have only considered a handful of attention variants, selected to reflect common module architectures for the respective tasks included in our analysis. Alternative attention specifications may yield different conclusions; and indeed we hope this work motivates further development of principled attention mechanisms (or encoders). Finally, we have limited our evaluation to tasks with unstructured output spaces, i.e., we have not considered seq2seq tasks, which we leave for future work. However we believe interpretability is more often a consideration in, e.g., classification than in translation.

Acknowledgements

References

David Alvarez-Melis and Tommi S Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. *arXiv preprint arXiv:1707.01943*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving machine attention from human rationales. *arXiv preprint arXiv:1808.09367*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512.

Shi Feng, Eric Wallace, Alvin Grissom II, Pedro Rodriguez, Mohit Iyyer, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretation difficult. In *Empirical Methods in Natural Language Processing*.

Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. 2018. Interpreting recurrent and attention-based neural models: a case study on natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4952–4957.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M Rush. 2017. Structured attention networks. *arXiv preprint arXiv:1702.00887*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

Tao Lei et al. 2017. *Interpretable neural models for natural language processing*. Ph.D. thesis, Massachusetts Institute of Technology.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Zachary C Lipton. 2016. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*, pages 1614–1623.

Azadeh Nikfarjam, Abeed Sarker, Karen OâĂŹConnor, Rachel Ginn, and Graciela Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3):671–681.

Nikolaos Pappas and Andrei Popescu-Belis. 2016. Human versus machine attention in document classification: A dataset with crowdsourced annotations. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 94–100.

Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.

Ben Peters, Vlad Niculae, and André FT Martins. 2018. Interpretable structure induction via sparse attention. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 365–367.

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on*

empirical methods in natural language processing, pages 1631–1642.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Qizhe Xie, Xuezhe Ma, Zihang Dai, and Eduard Hovy. 2017. An interpretable knowledge transfer model for knowledge base completion. *arXiv preprint arXiv:1704.05908*.

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Ye Zhang, Iain Marshall, and Byron C Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 2016, pages 795–804.

Appendices

A Model details

For all datasets, we use spaCy for tokenization. We map out of vocabulary words to a special <unk> token and map all words with numeric characters to 'qqq'. Each word in the vocabulary was initialized to pretrained embeddings. For general domain corpora we used either (i) FastText Embeddings (SST, IMDB, 20News, and CNN) trained on Simple English Wikipedia, or, (ii) GloVe 840B embeddings (AGNews and SNLI). For the MIMIC dataset, we learned word embeddings using Gensim over all discharge summaries in the corpus. We initialize words not present in the vocabulary using samples from a standard Gaussian $\mathcal{N}(\mu = 0, \sigma^2 = 1)$.

A.1 BiLSTM

We use an embedding size of 300 and hidden size of 128 for all datasets except bAbI (for which we use 50 and 30, respectively). All models were regularized using ℓ_2 regularization ($\lambda = 10^{-5}$) applied to all parameters. We use a sigmoid activation functions for binary classification tasks, and a softmax for all other outputs. We trained the model using maximum likelihood loss using the Adam Optimizer with default parameters in PyTorch.

A.2 CNN

We use an embedding size of 300 and 4 kernels of sizes [1, 3, 5, 7], each with 64 filters, giving a final hidden size of 256 (for bAbI we use 50 and 8 respectively with same kernel sizes). We use ReLU activation function on the output of the filters. All other configurations remain same as BiLSTM.

A.3 Average

We use the embedding size of 300 and a projection size of 256 with ReLU activation on the output of the projection matrix. All other configurations remain same as BiLSTM.

B Further details regarding attentional module of gradient

In the gradient experiments, we made the decision to cut-off the computation graph at the attention module so that gradient does not flow through this layer and contribute to the gradient feature importance score. For the sake of gradient calculation this effectively treats the attention as a separate input to the network, independent of the input. We argue that this is a natural choice to make for our analysis because it calculates: *how much does the output change as we perturb particular inputs (words) by a small amount, while paying the same amount of attention to said word as originally estimated and shown in the heatmap*?

C Correlations between Feature Importance measures

A question one might have here is how well correlated LOO and gradients are with *one another*. We report such results in their entirety on the paper website, and we summarize their correlations relative to those realized by attention in a BiLSTM model with LOO measures in Figure 6. This reports the mean differences between (i) gradient
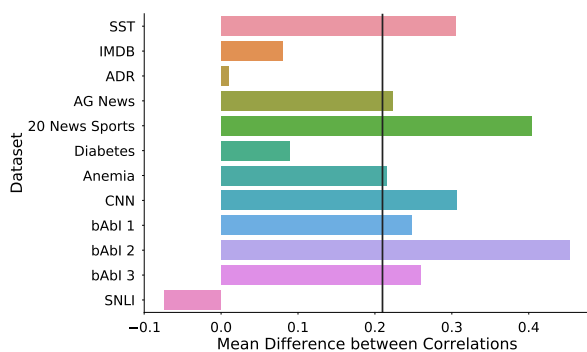
Figure 6: Mean difference in correlation of (i) LOO vs. Gradients and (ii) Attention vs. LOO scores using BiLSTM Encoder + Tanh Attention. On average the former is more correlated than the latter by >0.2 τ_{loo}.
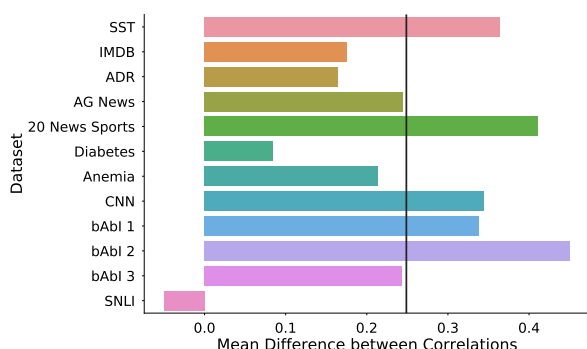

Figure 7: Mean difference in correlation of (i) LOO vs. Gradients and (ii) Attention vs. Gradients using BiLSTM Encoder + Tanh Attention. On average the former is more correlated than the latter by ~ 0.25 τ_g.

and LOO correlations, and (ii) attention and LOO correlations. As expected, we find that these exhibit, in general, considerably higher correlation with one another (on average) than LOO does with attention scores. (The lone exception is on SNLI.) Figure 7 shows the same for gradients and attention scores; the differences are comparable. In the *ADR* and *Diabetes* corpora, a few high precision tokens indicate (the positive) class, and in these cases we see better agreement between LOO/gradient measures with attention; this is consistent with Figure 4 which shows that it is difficult for the BiLSTM variant to find adversarial attention distributions for *Diabetes*.

A potential issues with using Kendall τ as our metric here is that (potentially many) irrelevant features may add noise to the correlation measures. We acknowledge that this as a shortcoming of the metric. One observation that may mitigate this concern is that we might expect such noise to depress the LOO and gradient correlations to the same extent as they do the correlation between at-
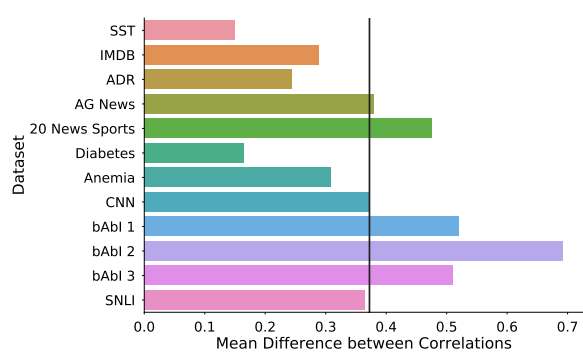

Figure 8: Difference in mean correlation of attention weights vs. LOO importance measures for (i) Average (feed-forward projection) and (ii) BiLSTM Encoders with Tanh attention. Average correlation (vertical bar) is on average ~ 0.375 points higher for the simple feedforward encoder, indicating greater correspondence with the LOO measure.

tention and feature importance scores; but as per Figure 7, they do not. We also note that the correlations between the attention weights on top of feedforward (projection) encoder and LOO scores are much stronger, on average, than those between BiLSTM attention weights and LOO. This is shown in Figure 8. Were low correlations due simply to noise, we would not expect this.[8]

D Graphs

To provide easy navigation of our (large set of) graphs depicting attention weights on various datasets/tasks under various model configuration we have created an interactive interface to browse these results, accessible at: https://successar.github.io/AttentionExplanation/docs/.

E Adversarial Heatmaps

SST

  Original: reggio falls victim to relying on the very digital technology that he fervently scorns creating a meandering inarticulate and ultimately disappointing film

  Adversarial: reggio falls victim to relying on the very digital technology that he fervently scorns creating a meandering inarticulate and ultimately disappointing film $\Delta\hat{y}$: *0.005*

IMDB

---

[8]The same contrast can be seen for the gradients, as one would expect given the direct gradient paths in the projection network back to individual tokens.

Original: fantastic movie one of the best film noir movies ever made bad guys bad girls a jewel heist a twisted morality a kidnapping everything is here jean has a face that would make bogart proud and the rest of the cast is is full of character actors who seem to to know they're onto something good get some popcorn and have a great time

Adversarial: fantastic movie one of the best film noir movies ever made bad guys bad girls a jewel heist a twisted morality a kidnapping everything is here jean has a face that would make bogart proud and the rest of the cast is is full of character actors who seem to to know they're onto something good get some popcorn and have a great time $\triangle\hat{y}$: *0.004*

20 News Group - Sports

Original: i meant to comment on this at the time there ' s just no way baserunning could be that important if it was runs created would n ' t be nearly as accurate as it is runs created is usually about qqq qqq accurate on a team level and there ' s a lot more than baserunning that has to account for the remaining percent .

Adversarial: i meant to comment on this at the time there ' s just no way baserunning could be that important if it was runs created would n ' t be nearly as accurate as it is runs created is usually about qqq qqq accurate on a team level and there ' s a lot more than baserunning that has to account for the remaining percent . $\triangle\hat{y}$: *0.001*

ADR

Original: meanwhile wait for DRUG and DRUG to kick in first co i need to prep dog food etc . co omg <UNK> .

Adversarial: meanwhile wait for DRUG and DRUG to kick in first co i need to prep dog food etc . co omg <UNK> . $\triangle\hat{y}$: *0.002*

AG News

Original: general motors and daimlerchrysler say they # qqq teaming up to develop hybrid technology for use in their vehicles . the two giant automakers say they have signed a memorandum of understanding

Adversarial: general motors and daimlerchrysler say they # qqq teaming up to develop hybrid technology for use in their vehicles . the two giant automakers say they have signed a memorandum of understanding . $\triangle\hat{y}$: *0.006*

SNLI

Hypothesis: a man is running on foot

Original Premise Attention: a man in a gray shirt and blue shorts is standing outside of an old fashioned ice cream shop named sara 's old fashioned ice cream , holding his bike up , with a wood like table , chairs , benches in front of him .

Adversarial Premise Attention: a man in a gray shirt and blue shorts is standing outside of an old fashioned ice cream shop named sara 's old fashioned ice cream , holding his bike up , with a wood like table , chairs , benches in front of him . $\triangle\hat{y}$: *0.002*

Babi Task 1

Question: Where is Sandra ?

Original Attention: John travelled to the garden . Sandra travelled to the garden

Adversarial Attention: John travelled to the garden . Sandra travelled to the garden $\triangle\hat{y}$: *0.003*

CNN-QA

Question: federal education minister @placeholder visited a @entity15 store in @entity17 , saw cameras

Original: @entity1 , @entity2 (@entity3) police have arrested four employees of a popular @entity2 ethnic - wear chain after a minister spotted a security camera overlooking the changing room of one of its stores . federal education minister @entity13 was visiting a @entity15 outlet in the tourist resort state of @entity17 on friday when she discovered a surveillance camera pointed at the changing room , police said . four employees of the store have been arrested , but its manager – herself a woman – was still at large saturday , said @entity17 police superintendent @entity25 . state authorities launched their investigation right after @entity13 levied her accusation . they found an overhead camera that the minister had spotted and determined that it was indeed able to take photos of customers using the store 's changing room , according to @entity25 . after the incident , authorities sealed off the store and summoned six top officials from @entity15 , he said . the arrested staff have been charged with voyeurism and breach of privacy , according to the police . if convicted , they could spend up to three years in jail , @entity25 said . officials from @entity15 – which sells ethnic garments , fabrics and other products – are heading to @entity17 to work

with investigators , according to the company .
" @entity15 is deeply concerned and shocked at
this allegation , " the company said in a statement
. " we are in the process of investigating this
internally and will be cooperating fully with the
police . "

Adversarial: @entity1 , @entity2 (@entity3
) police have arrested four employees of a
popular @entity2 ethnic - wear chain after a
minister spotted a security camera overlooking
the changing room of one of its stores . federal
education minister @entity13 was visiting a
@entity15 outlet in the tourist resort state of
@entity17 on friday when she discovered a
surveillance camera pointed at the changing room
, police said . four employees of the store
have been arrested , but its manager – herself
a woman – was still at large saturday , said
@entity17 police superintendent @entity25 . state
authorities launched their investigation right after
@entity13 levied her accusation . they found an
overhead camera that the minister had spotted
and determined that it was indeed able to take
photos of customers using the store 's changing
room , according to @entity25 . after the incident
, authorities sealed off the store and summoned
six top officials from @entity15 , he said . the
arrested staff have been charged with voyeurism
and breach of privacy , according to the police .
if convicted , they could spend up to three years
in jail , @entity25 said . officials from @entity15
– which sells ethnic garments , fabrics and other
products – are heading to @entity17 to work
with investigators , according to the company .
" @entity15 is deeply concerned and shocked at
this allegation , " the company said in a statement
. " we are in the process of investigating this
internally and will be cooperating fully with the
police . " $\Delta \hat{y}$: *0.005*