Attention is not not Explanation

Sarah Wiegreffe*
School of Interactive Computing
Georgia Institute of Technology
saw@gatech.edu

Yuval Pinter*
School of Interactive Computing
Georgia Institute of Technology
uvp@gatech.edu

Abstract

Attention mechanisms play a central role in NLP systems, especially within recurrent neural network (RNN) models. Recently, there has been increasing interest in whether or not the intermediate representations offered by these modules may be used to explain the reasoning for a model's prediction, and consequently reach insights regarding the model's decision-making process. A recent paper claims that 'Attention is not Explanation' (Jain and Wallace, 2019). We challenge many of the assumptions underlying this work, arguing that such a claim depends on one's definition of explanation, and that testing it needs to take into account all elements of the model. We propose four alternative tests to determine when/whether attention can be used as explanation: a simple uniform-weights baseline; a variance calibration based on multiple random seed runs; a diagnostic framework using frozen weights from pretrained models; and an end-to-end adversarial attention training protocol. Each allows for meaningful interpretation of attention mechanisms in RNN models. We show that even when reliable adversarial distributions can be found, they don't perform well on the simple diagnostic, indicating that prior work does not disprove the usefulness of attention mechanisms for explainability.

1 Introduction

Attention mechanisms (Bahdanau et al., 2014) are nowadays ubiquitous in NLP, and their suitability for providing explanations for model predictions is a topic of high interest (Xu et al., 2015; Rocktäschel et al., 2015; Mullenbach et al., 2018; Thorne et al., 2019; Serrano and Smith, 2019). If they indeed offer such insights, many application areas would benefit by better understanding the internals of neural models that use attention as a means for, e.g., model debugging or architecture selection. A recent paper (Jain and Wallace, 2019) points to possible pitfalls that may cause researchers to misapply attention scores as explanations of model behavior, based on a premise that explainable attention distributions should be *consistent* with other feature-importance measures as well as *exclusive* given a prediction.[1] Its core argument, which we elaborate in §2, is that if alternative attention distributions exist that produce similar results to those obtained by the original model, then the original model's attention scores cannot be reliably used to "faithfully" explain the model's prediction. Empirically, the authors show that achieving such alternative distributions is easy for a large sample of English-language datasets.

Idea of Jain and Wallace paper

We contend (§2.1) that while Jain and Wallace ask an important question, and raise a genuine concern regarding potential misuse of attention weights in explaining model decisions on English-language datasets, some key assumptions used in their experimental design leave an implausibly large amount of freedom in the setup, ultimately leaving practitioners without an applicable way for measuring the utility of attention distributions in specific settings.

We apply a more model-driven approach to this question, beginning (§3.2) with testing attention modules' **contribution** to a model by applying a simple baseline where attention weights are frozen to a uniform distribution. We demonstrate that for some datasets, a frozen attention distribution performs just as well as learned attention weights, concluding that randomly- or adversarially-perturbed distributions are not ev-

*Equal contributions.

[1] A preliminary version of our theoretical argumentation was published as a blog post on Medium at http://bit.ly/2OTzU4r. Following the ensuing online discussion, the authors uploaded a post-conference version of the paper to arXiv (v3) which addresses some of the issues in the post. We henceforth refer to this later version.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 11–20, Hong Kong, China, November 3–7, 2019.

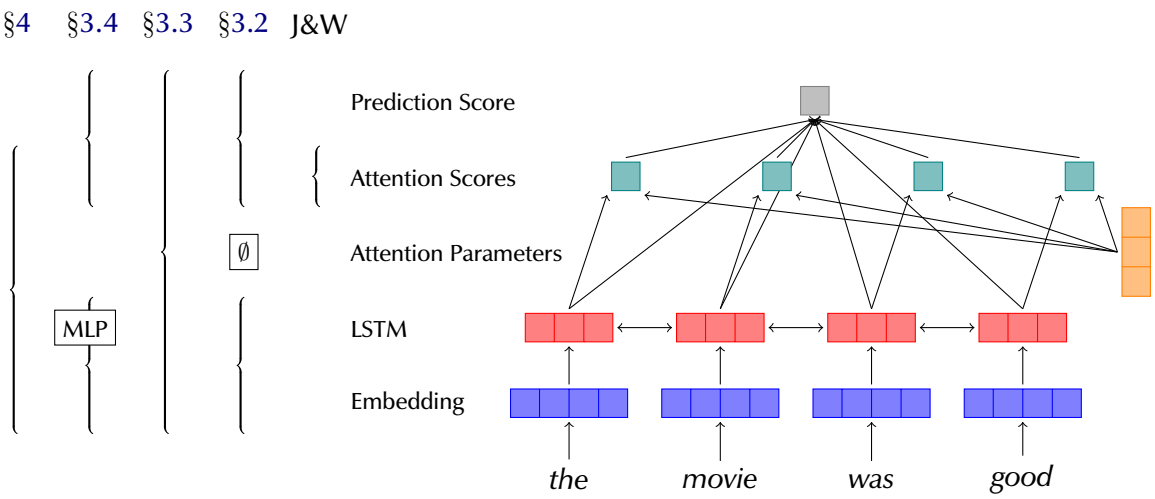


Figure 1: Schematic diagram of a classification LSTM model with attention, including the components manipulated or replaced in the experiments performed in Jain and Wallace (2019) and in this work (by section).

idence against attention as explanation in these cases. We next (§3.3) examine the **expected variance** in attention-produced weights by initializing multiple training sequences with different random seeds, allowing a better quantification of how much variance can be expected in trained models. We show that considering this background stochastic variation when comparing adversarial results with a traditional model allows us to better interpret adversarial results. In §3.4, we present a simple yet effective **diagnostic tool** which tests attention distributions for their usefulness by using them as frozen weights in a non-contextual multi-layered perceptron (MLP) architecture. The favorable performance of LSTM-trained weights provides additional support for the coherence of trained attention scores. This demonstrates a sense in which attention components indeed provide a meaningful model-agnostic interpretation of tokens in an instance.

In §4, we introduce a **model-consistent** training protocol for finding adversarial attention weights, correcting some flaws we found in the previous approach. We train a model using a modified loss function which takes into account the distance from an ordinarily-trained base model's attention scores in order to learn parameters for adversarial attention distributions. We believe these experiments are now able to support or refute a claim of faithful explainability, by providing a way for convincingly saying by construction that a plausible alternative 'explanation' can (or cannot) be constructed for a given dataset and model architecture. We find that while plausibly adversarial distributions of the consistent kind can indeed be found for the binary classification datasets in question, they are not as extreme as those found in the inconsistent manner, as illustrated by an example from the IMDB task in Figure 2. Furthermore, these outputs do not fare well in the diagnostic MLP, calling into question the extent to which we can treat them as equally powerful for explainability.

Finally, we provide a theoretical discussion (§5) on the definitions of interpretability and explainability, grounding our findings within the accepted definitions of these concepts.

Our four quantitative experiments are illustrated in Figure 1, where each bracket on the left covers the components in a standard RNN-with-attention architecture which we manipulate in each experiment. We urge NLP researchers to consider applying the techniques presented here on their models containing attention in order to evaluate its effectiveness at providing explanation. We offer our code for this purpose at `https://github.com/sarahwie/attention`.

2 Attention Might be Explanation

In this section, we briefly describe the experimental design of Jain and Wallace (2019) and look at the results they provide to support their claim that 'Attention is not explanation'. The authors select eight classification datasets, mostly binary, and two question answering tasks for their experiments (detailed in §3.1).

They first present a correlation analysis of attention scores and other interpretability measures.

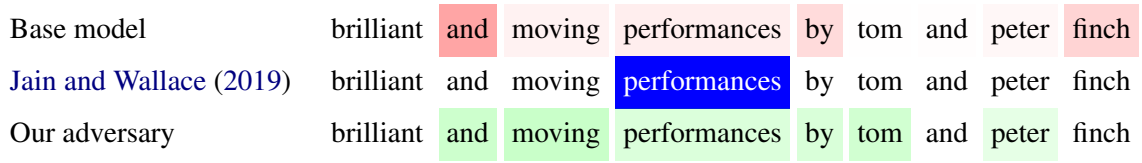


Figure 2: Attention maps for an IMDb instance (all predicted as positive with score > 0.998), showing that in practice it is difficult to learn a distant adversary which is consistent on all instances in the training set.

They find that attention is not strongly correlated with other, well-grounded feature importance metrics, specifically gradient-based and leave-one-out methods (which in turn correlate well with each other). This experiment evaluates the authors' claim of *consistency* – that attention-based methods of explainability cannot be valid if they do not correlate well with other metrics. We find the experiments in this part of the paper convincing and do not focus our analysis here. We offer our simple MLP diagnostic network (§3.4) as an additional way for determining validity of attention distributions, in a more *in vivo* setting.

Next, the authors present an adversarial search for alternative attention distributions which minimally change model predictions. To this end, they manipulate the attention distributions of trained models (which we will call **base** from now on) to discern whether alternative distributions exist for which the model outputs near-identical prediction scores. They are able to find such distributions, first by randomly permuting the base attention distributions on the test data during model inference, and later by adversarially searching for maximally different distributions that still produce a prediction score within ϵ of the base distribution. They use these experimental results as supporting evidence for the claim that attention distributions cannot be explainable because they are not *exclusive*. As stated, the lack of comparable change in prediction with a change in attention scores is taken as evidence for a lack of "faithful" explainability of the attention mechanism from inputs to output.

Notably, Jain and Wallace detach the attention distribution and output layer of their pretrained network from the parameters that compute them (see Figure 1), treating each attention score as a standalone unit independent of the model. In addition, they compute an independent adversarial distribution for each instance.

2.1 Main Claim

We argue that Jain and Wallace's counterfactual attention weight experiments do not advance their thesis, for the following reasons:

Attention Distribution is not a Primitive. From a modeling perspective, detaching the attention scores obtained by parts of the model (i.e. the attention mechanism) degrades the model itself. The base attention weights are not assigned arbitrarily by the model, but rather computed by an integral component whose parameters were trained alongside the rest of the layers; the way they work depends on each other. Jain and Wallace provide alternative distributions which may result in similar predictions, but in the process they remove the very linkage which motivates the original claim of attention distribution explainability, namely the fact that the model was *trained* to attend to the tokens it chose. A reliable adversary must take this consideration into account, as our setup in §4 does.

Existence does not Entail Exclusivity. On a more theoretical level, we hold that attention scores are used as providing *an* explanation; not *the* explanation. The final layer of an LSTM model may easily produce outputs capable of being aggregated into the same prediction in various ways, however the model still makes the choice of a specific weighting distribution using its trained attention component. This mathematically flexible production capacity is particularly evident in binary classifiers, where prediction is reduced to a single scalar, and an average instance (of e.g. the IMDB dataset) might contain 179 tokens, i.e. 179 scalars to be aggregated. This effect is greatly exacerbated when performed independently on each instance.[2] Thus, it is no surprise that Jain and Wal-

[2]Indeed, the most open-ended task, question answering over CNN data, produces considerable difficulty to manipulate its scores by random permutation (Figure 6e in Jain and Wallace (2019)). Similarly, the adversarial examples presented in Appendix C of the paper for the QA datasets select a different token of the correct word's type, which should not surprise us even under an LSTM assumption (encoder hidden

Dataset	Avg. Length (tokens)	Train Size (neg/pos)	Test Size (neg/pos)
Diabetes	1858	6381/1353	1295/319
Anemia	2188	1847/3251	460/802
IMDb	179	12500/12500	2184/2172
SST	19	3034/3321	863/862
AgNews	36	30000/30000	1900/1900
20News	115	716/710	151/183

Table 1: Dataset statistics.

Dataset	Attention (Base)		Uniform
	Reported	Reproduced	
Diabetes	0.79	0.775	0.706
Anemia	0.92	0.938	0.899
IMDb	0.88	0.902	0.879
SST	0.81	0.831	0.822
AgNews	0.96	0.964	0.960
20News	0.94	0.942	0.934

Table 2: Classification F1 scores (1-class) on attention models, both as reported by Jain and Wallace and in our reproduction, and on models forced to use uniform attention over hidden states.

lace find what they are looking for given this degree of freedom.

In summary, due to the per-instance nature of the demonstration and the fact that model parameters have not been learned or manipulated directly, Jain and Wallace have **not shown the existence of an adversarial model** that produces the claimed adversarial distributions. Thus, we cannot treat these adversarial attentions as equally plausible or faithful explanations for model prediction. Additionally, they haven't provided a baseline of how much variation is to be expected in learned attention distributions, leaving the reader to question just how adversarial the found adversarial distributions are.

3 Examining Attention Distributions

In this section, we apply a careful methodological approach for examining the properties of attention distributions and propose alternatives. We begin by identifying the appropriate scope of the models' performance and variance, followed by implementing an empirical diagnostic technique which measures the model-agnostic usefulness of attention weights in capturing the relationship between inputs and output.

3.1 Experimental Setup

In order to make our many points in a succinct fashion as well as follow the conclusions drawn by Jain and Wallace, we focus on experimenting with the binary classification subset of their tasks, and on models with an LSTM architecture (Hochreiter and Schmidhuber, 1997), the only one the authors make firm conclusions on. Future work may extend our experiments to extractive tasks like question answering, as well as other attention-prone tasks, like seq2seq models.

We experiment on the following datasets: Stanford Sentiment Treebank (SST) (Socher et al., 2013), IMDB Large Movie Reviews Corpus (Maas et al., 2011), 20 NEWSGROUPS (hockey vs. baseball),[3] the AG NEWS Corpus,[4] and two prediction tasks from MIMIC-III ICD9 (Johnson et al., 2016): DIABETES and ANEMIA. The tasks are as follows: to predict positive or negative sentiment from sentences (SST) and movie reviews (IMDB), to predict the topic of news articles as either baseball (neg.) or hockey (pos.) in 20 NEWSGROUPS and either world (neg.) or business (pos.) in AG NEWS, to predict whether a patient is diagnosed with diabetes from their ICU discharge summary, and to predict whether the patient is diagnosed with acute (neg.) or chronic (pos.) anemia (both MIMIC-III ICD9). We use the dataset versions, including train-test split, provided by Jain and Wallace.[5] All datasets are in English.[6] Data statistics are provided in Table 1.

We use a single-layer bidirectional LSTM with tanh activation, followed by an additive attention layer (Bahdanau et al., 2014) and softmax prediction, which is equivalent to the LSTM setup of Jain and Wallace. We use the same hyperparameters found in that work to be effective in training, which we corroborated by reproducing its results to a satisfactory degree (see middle columns of Table 2). We refer to this architecture as the **main setup**, where training results in a **base model**.

Following Jain and Wallace, all analysis is performed on the test set. We report F1 scores on the positive class, and apply the same metrics they use for model comparison, namely Total Variation

states are typically affected by the input word to a noticeable degree).

[3] http://qwone.com/~jason/20Newsgroups/

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

[5] https://github.com/successar/AttentionExplanation

[6] We do not include the Twitter Adverse Drug Reactions (ADR) (Nikfarjam et al., 2015) dataset as the source tweets are no longer all available.

Distance (TVD) for comparing prediction scores $\hat{y}$ and Jensen-Shannon Divergence (JSD) for comparing weighting distributions α:

$$\text{TVD}(\hat{y}_1, \hat{y}_2) = \frac{1}{2} \sum_{i=1}^{|\mathcal{Y}|} |\hat{y}_{1i} - \hat{y}_{2i}| \,;$$

$$\text{JSD}(\alpha_1, \alpha_2) = \frac{1}{2} \text{KL}[\alpha_1 \parallel \bar{\alpha}] + \frac{1}{2} \text{KL}[\alpha_2 \parallel \bar{\alpha}],$$

where $\bar{\alpha} = \frac{\alpha_1 + \alpha_2}{2}$.

3.2 Uniform as the Adversary

First, we test the validity of the classification tasks and datasets by examining whether attention is necessary in the first place. We argue that if attention models are not useful compared to very simple baselines, i.e. their parameter capacity is not being used, there is no point in using their outcomes for any type of explanation to begin with. We thus introduce a **uniform** model variant, identical to the main setup except that the attention distribution is frozen to uniform weights over the hidden states.

The results comparing this baseline with the base model are presented in Table 2. If attention was a necessary component for good performance, we would expect a large drop between the two rightmost columns. Somewhat surprisingly, for three of the classification tasks the attention layer appears to offer little to no improvement whatsoever. We conclude that these datasets, notably AG NEWS and 20 NEWSGROUPS, are not useful test cases for the debated question: attention is not explanation if you don't need it. We subsequently ignore the two News datasets, but keep SST, which we deem borderline.

3.3 Variance within a Model

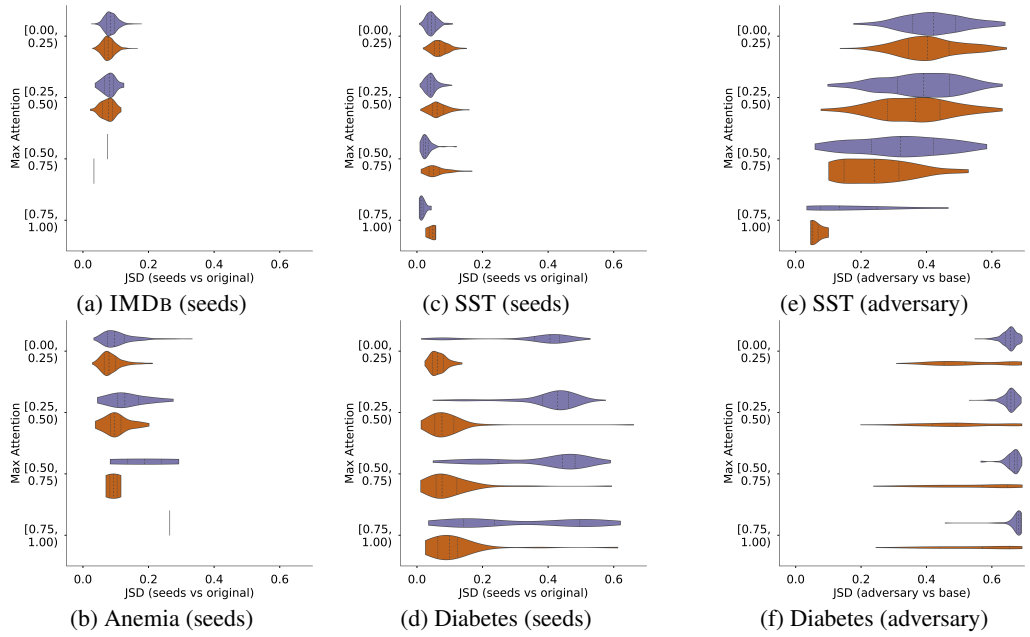
We now test whether the variances observed by Jain and Wallace between trained attention scores and adversarially-obtained ones are unusual. We do this by repeating their analysis on eight models trained from the main setup using different initialization random seeds. The variance introduced in the attention distributions represents a baseline amount of variance that would be considered normal.

The results are plotted in Figure 3 using the same plane as Jain and Wallace's Figure 8 (with two of these reproduced as (e-f)). Left-heavy violins are interpreted as data classes for which the compared model produces attention distributions similar to the base model, and so having an adversary that manages to 'pull right' supports the argument that distributions are easy to manipulate. We see that SST distributions (c, e) are surprisingly robust to random seed change, validating our choice to continue examining this dataset despite its borderline F1 score. On the Diabetes dataset, the negative class is already subject to relatively arbitrary distributions from the different random seed settings (d), making the highly divergent results from the overly-flexible adversarial setup (f) seem less impressive. Our consistently-adversarial setup in §4 will further explore the difficulty of surpassing seed-induced variance between attention distributions.

3.4 Diagnosing Attention Distributions by Guiding Simpler Models

As a more direct examination of models, and as a complementary approach to Jain and Wallace (2019)'s measurement of backward-pass gradient flows through the model for gauging token importance, we introduce a post-hoc training protocol of a non-contextual model **guided** by pre-set weight distributions. The idea is to examine the prediction power of attention distributions in a 'clean' setting, where the trained parts of the model have no access to neighboring tokens of the instance. If pre-trained scores from an attention model perform well, we take this to mean they are helpful and consistent, fulfilling a certain sense of explainability. In addition, this setup serves as an effective diagnostic tool for assessing the utility of adversarial attention distributions: if such distributions are truly alternative, they should be equally useful as guides as their base equivalent, and thus perform comparably.

Our diagnostic model is created by replacing the main setup's LSTM and attention parameters with a token-level affine hidden layer with $\tanh$ activation (forming an MLP), and forcing its output scores to be weighted by a pre-set, per-instance distribution, during both training and testing. This setup is illustrated in Figure 4. The guide weights we impose are the following: **Uniform**, where we force the MLP outputs to be considered equally across each instance, effectively forming an unweighted baseline; **Trained MLP**, where we do not freeze the weights layer, instead allowing the



(a) IMDB (seeds) (c) SST (seeds) (e) SST (adversary)

(b) Anemia (seeds) (d) Diabetes (seeds) (f) Diabetes (adversary)

Figure 3: Densities of maximum JS divergences (x-axis) as a function of the max attention (y-axis) in each instance between the base distributions and: (a-d) models initialized on different random seeds; (e-f) models from a per-instance adversarial setup (replication of Figure 8a, 8c resp. in Jain and Wallace (2019)). In each max-attention bin, top (blue) is the negative-label instances, bottom (red) positive-label instances.

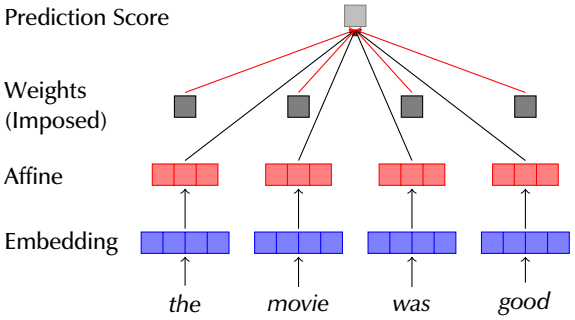


Figure 4: Diagram of the setup in §3.4 (except TRAINED MLP, which learns weight parameters).

Guide weights	Diab.	Anemia	SST	IMDb
UNIFORM	0.404	0.873	0.812	0.863
TRAINED MLP	0.699	0.920	0.817	0.888
BASE LSTM	**0.753**	0.931	**0.824**	**0.905**
ADVERSARY (4)	0.503	**0.932**	0.592	0.700

Table 3: F1 scores on the positive class for an MLP model trained on various weighting guides. For ADVERSARY, we set $\lambda \leftarrow 0.001$.

MLP to learn its own attention parameters;[7] **Base LSTM**, where we take the weights learned by the base LSTM model's attention layer; and **Adversary**, based on distributions found adversarially using the consistent training algorithm from §4 below (where their results will be discussed).

The results are presented in Table 3. The first important result, consistent across datasets, is that using pre-trained LSTM attention weights is better than letting the MLP learn them on its own, which is in turn better than the unweighted baseline. Comparing with results from §3.2, we see that this setup also outperforms the LSTM trained with uniform attention weights, suggesting that the attention module is more important than the word-level architecture for these datasets. These findings strengthen the case counter to the claim that attention weights are arbitrary: independent token-level models that have no access to contextual information find them useful, indicating that they encode some measure of token importance which is not model-dependent.

4 Training an Adversary

Having demonstrated three methods which test the meaningfulness of attention distributions as instruments of explainability with adequate control, we now propose a model-consistent training protocol for finding adversarial attention distributions through a coherent parameterization, which holds across all training instances. We believe this setup is able to advance the search for faithful explainability (see §5). Indeed, our results

[7] This is the same as Jain and Wallace's *average* setup.

will demonstrate that the extent to which a model-consistent adversary can be found varies across datasets, and that the dramatic reduction in degree of freedom compared to previous work allows for better-informed analysis.

Model. Given the base model $\mathcal{M}_\text{b}$, we train a model $\mathcal{M}_\text{a}$ whose explicit goal is to provide similar prediction scores for each instance, while distancing its attention distributions from those of $\mathcal{M}_\text{b}$. Formally, we train the adversarial model using stochastic gradient updates based on the following loss formula (summed over instances in the minibatch):

$$\mathcal{L}(\mathcal{M}_a, \mathcal{M}_b)^{(i)} = \text{TVD}(\hat{y}_a^{(i)}, \hat{y}_b^{(i)}) - \lambda \ \text{KL}(\boldsymbol{\alpha}_a^{(i)} \parallel \boldsymbol{\alpha}_b^{(i)}),$$

where $\hat{y}^{(i)}$ and $\boldsymbol{\alpha}^{(i)}$ denote predictions and attention distributions for an instance i, respectively.

λ is a hyperparameter which we use to control the tradeoff between relaxing the prediction distance requirement (low TVD) in favor of more divergent attention distributions (high JSD), and vice versa. When this interaction is plotted on a two-dimensional axis, the shape of the plot can be interpreted to either support the 'attention is not explanation' hypothesis if it is convex (JSD is easily manipulable), or oppose it if it is concave (early increase in JSD comes at a high cost in prediction precision).

Prediction performance. By definition, our loss objective does not directly consider actual prediction performance. The TVD component pushes it towards the same score as the base model, but our setup does not ensure generalization from train to test. It would thus be interesting to inspect the extent of the implicit F1/TVD relationship. We report the highest F1 scores of models whose attention distributions diverge from the base, on average, by at least 0.4 in JSD, as well as their λ setting and corresponding comparison metrics, in Table 4 (full results available in Appendix B). All F1 scores are on par with the original model results reported in Table 2, indicating the effectiveness of our adversarial models at imitating base model scores on the test sets.

Adversarial weights as guides. We next apply the diagnostic setup introduced in §3.4 by training a guided MLP model on the adversarially-trained attention distributions. The results, reported in the bottom line of Table 3, show that despite their local decision-imitation abilities, they are usually completely incapable of providing a non-contextual framework with useful guides.[8] We offer these results as evidence that adversarial distributions, even those obtained consistently for a dataset, deprive the underlying model from some form of understanding it gained over the data, one that it was able to leverage by tuning the attention mechanism towards preferring 'useful' tokens.

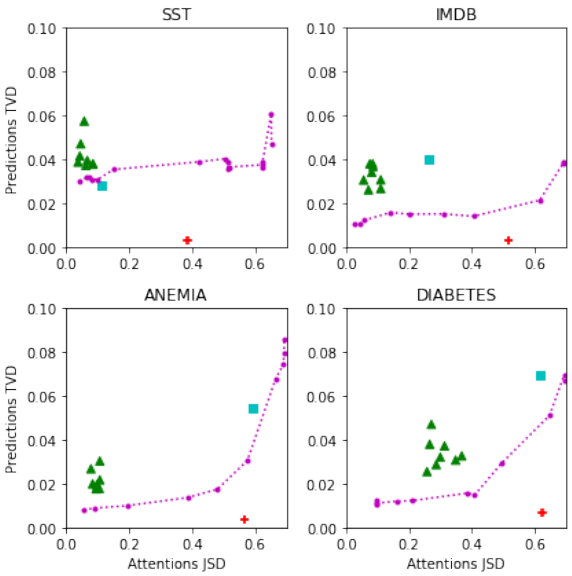


Figure 5: Averaged per-instance test set JSD and TVD from base model for each model variant. JSD is bounded at $\sim$ 0.693. ▲: random seed; ■: uniform weights; dotted line: our adversarial setup as λ is varied; +: adversarial setup from Jain and Wallace (2019).

Dataset	λ	F1 (↑)	TVD (↓)	JSD (↑)
Diabetes	2e-4	0.775	0.015	0.409
Anemia	5e-4	0.942	0.017	0.481
SST	5.25e-4	0.823	0.036	0.514
IMDb	8e-4	0.906	0.014	0.405

Table 4: Best-performing adversarial models with instance-average JSD > 0.4.

TVD/JSD tradeoff. In Figure 5 we present the levels of prediction variance (TVD) allowed

[8] We note the outlying result achieved on the Anemia dataset. This can be explained via the data distribution, which is heavily skewed towards positive examples (see Table 1 in the Appendix), together with the fact (conceded in Jain and Wallace (2019)'s section 4.2.1) that positive instances in detection datasets such as MIMIC tend to contain a handful of indicative tokens, making the particularly helpful distributions reached by a trained model hard to replace by an adversary. Together, this leads to the selected setting of $\lambda = 0.001$ producing average distributions substantially more similar to the base than in the other datasets (JSD $\sim$ 0.58 vs. $>$ 0.61) and thus more useful to the MLP setup.

by models achieving increased attention distance (JSD) on all four datasets. The convex shape of most curves does lend support to the claim that attention scores are easily manipulable; however the extent of this effect emerging from Jain and Wallace's per-instance setup is a considerable exaggeration, as seen by its position (+) well below the curve of our parameterized model set. Again, the SST dataset emerges as an outlier: not only can JSD be increased practically arbitrarily without incurring prediction variance cost, the uniform baseline (■) comes up under the curve, i.e. with a better adversarial score. We again include random seed initializations (▲) in order to quantify a baseline amount of variance.

TVD/JSD plots broken down by prediction class are available in Appendix C. In future work, we intend to inspect the potential of multiple adversarial attention models existing side-by-side, all distant enough from each other.

Concrete Example. Table 2 illustrates the difference between inconsistently-achieved adversarial heatmaps and consistently trained ones. Despite both adversaries approximating the desired prediction score to very high degree, the heatmaps show that Jain and Wallace's model has distributed all of the attention weight to an ad-hoc token, whereas our trained model could only distance itself from the base model distribution by so much, keeping multiple tokens in the > 0.1 score range.

5 Defining Explanation

The umbrella term of "Explainable AI" encompasses at least three distinct notions: *transparency*, *explainability*, and *interpretability*. Lipton (2016) categorizes transparency, or overall human understanding of a model, and post-hoc explainability as two competing notions under the umbrella of interpretability. The relevant sense of transparency, as defined by Lipton (2016) (§3.1.2), pertains to the way in which a specific portion of a model corresponds to a human-understandable construct (which Doshi-Velez and Kim (2017) refer to as a "cognitive chunk"). Under this definition, it should appear sensible of the NLP community to treat attention scores as a vehicle of (partial) transparency. Attention mechanisms do provide a look into the inner workings of a model, as they produce an easily-understandable weighting of hidden states.

Rudin (2018) defines explainability as simply a plausible (but not necessarily faithful) reconstruction of the decision-making process, and Riedl (2019) classifies explainable rationales as valuable in that they mimic what we as humans do when we rationalize past actions: we invent a story that plausibly justifies our actions, even if it is not an entirely accurate reconstruction of the neural processes that produced our behavior at the time. Distinguishing between interpretability and explainability as two separate notions, Rudin (2018) argues that interpretability is more desirable but more difficult to achieve than explainability, because it requires presenting humans with a big-picture understanding of the correlative relationship between inputs and outputs (citing the example of linear regression coefficients). Doshi-Velez and Kim (2017) break down interpretability into further subcategories, depending on the amount of human involvement and the difficulty of the task.

In prior work, Lei et al. (2016) train a model to simultaneously generate rationales and predictions from input text, using gold-label rationales to evaluate their model. Generally, many accept the notion of extractive methods such as Lei et al. (2016), in which explanations come directly from the input itself (as in attention), as plausible. Works such as Mullenbach et al. (2018) and Ehsan et al. (2019) use human evaluation to evaluate explanations; the former based on attention scores over the input, and the latter based on systems with additional rationale-generation capability. The authors show that rationales generated in a post-hoc manner increase user trust in a system.

Citing Ross et al. (2017), Jain and Wallace's requisite for attention distributions to be used as explanation is that there must only exist one or a few closely-related correct explanations for a model prediction. However, Doshi-Velez and Kim (2017) caution against applying evaluations and terminology broadly without clarifying task-specific explanation needs. If we accept the Rudin and Riedl definitions of explainability as providing a *plausible*, but not necessarily *faithful* rationale for model prediction, then the argument against attention mechanisms because they are not exclusive as claimed by Jain and Wallace is invalid, and human evaluation (which they do not consult) is necessary to evaluate the plausibility of generated rationales. Just because there exists another explanation does not mean that the one provided is false

or meaningless, and under this definition the existence of multiple different explanations is not necessarily indicative of the quality of a single one.

Jain and Wallace define attention and explanation as measuring the "responsibility" each input token has on a prediction. This aligns more closely with the more rigorous (Lipton, 2016, §3.1.1) definition of transparency, or Rudin (2018)'s definition of interpretability: human understanding of the model as a whole rather than of its respective parts. The ultimate question posed so far as 'is attention explanation?' seems to be: do high attention weights on certain elements in the input lead the model to make its prediction? This question is ultimately left largely unanswered by prior work, as we address in previous sections. However, under the given definition of transparency, the authors' exclusivity requisite is well-defined and we find value in their counterfactual framework as a concept – if a model is capable of producing multiple sets of diverse attention weights for the same prediction, then the relationship between inputs and outputs used to make predictions is not understood by attention analysis. This provides us with the motivation to implement the adversarial setup coherently and to derive and present conclusions from it. To this end, we additionally provide our §3.4 model to test the relationship between input tokens and output.

In the terminology of Doshi-Velez and Kim (2017), our proposed methods provide a *functionally-grounded* evaluation of attention as explanation, i.e. an analysis conducted on proxy tasks without human evaluation. We believe the proxies we have provided can be used to test the validity of attention as a form of explanation from the ground-up, based on the type of explanation one is looking for.

6 Attention is All you Need it to Be

Whether or not attention is explanation depends on the definition of explainability one is looking for: *plausible* or *faithful* explanations (or both). We believe that prior work focused on providing plausible rationales is not invalidated by Jain and Wallace's or our results. However, we have confirmed that adversarial distributions can be found for LSTM models in some classification tasks, as originally hypothesized by Jain and Wallace. This should provide pause to researchers who are looking to attention distributions for one true, faithful interpretation of the link their model has established between inputs and outputs. At the same time, we have provided a suite of experiments that researchers can make use of in order to make informed decisions about the quality of their models' attention mechanisms when used as explanation for model predictions.

We've shown that alternative attention distributions found via adversarial training methods perform poorly relative to traditional attention mechanisms when used in our diagnostic MLP model. These results indicate that trained attention mechanisms in RNNs on our datasets do in fact learn something meaningful about the relationship between tokens and prediction which cannot be easily 'hacked' adversarially.

We view the conditions under which adversarial distributions can actually be found in practice to be an important direction for future work. Additional future directions for this line of work include application on other tasks such as sequence modeling and multi-document analysis (NLI, QA); extension to languages other than English; and adding a human evaluation for examining the level of agreement with our measures. We also believe our work can provide value to theoretical analysis of attention models, motivating development of analytical methods to estimate the usefulness of attention as an explanation based on dataset and model properties.

Acknowledgments

We thank Yoav Goldberg for preliminary comments on the idea behind the original Medium post. We thank the online community who participated in the discussion following the post, and particularly Sarthak Jain and Byron Wallace for their active engagement, as well as for the high-quality code they released which allowed fast reproduction and modification of their experiments. We thank Erik Wijmans for early feedback. We thank the members of the Computational Linguistics group at Georgia Tech for discussions and comments, particularly Jacob Eisenstein and Murali Raghu Babu. We thank the anonymous reviewers for many useful comments.

YP is a Bloomberg Data Science PhD Fellow.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly

learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 263–274. ACM.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

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-volume 1*, pages 142–150. Association for Computational Linguistics.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Azadeh Nikfarjam, Abeed Sarker, Karen O'connor, 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.

Mark O Riedl. 2019. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1):33–36.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2662–2670. AAAI Press.

Cynthia Rudin. 2018. Please stop explaining black box models for high stakes decisions. *arXiv preprint arXiv:1811.10154*.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

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.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2019. Generating token-level explanations for natural language inference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 963–969, Minneapolis, Minnesota. Association for Computational Linguistics.

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.