Abstract

Attention mechanisms are ubiquitous components in neural architectures applied to natural language processing. In addition to yielding gains in predictive accuracy, attention weights are often claimed to confer interpretability, purportedly useful both for providing insights to practitioners and for explaining why a model makes its decisions to stakeholders. We call the latter use of attention mechanisms into question by demonstrating a simple method for training models to produce deceptive attention masks. Our method diminishes the total weight assigned to designated impermissible tokens, even when the models can be shown to nevertheless rely on these features to drive predictions. Across multiple models and tasks, our approach manipulates attention weights while paying surprisingly little cost in accuracy. Through a human study, we show that our manipulated attention-based explanations deceive people into thinking that predictions from a model biased against gender minorities do not rely on the gender. Consequently, our results cast doubt on attention’s reliability as a tool for auditing algorithms in the context of fairness and accountability.

1 Introduction

Since their introduction as a method for aligning inputs and outputs in neural machine translation, attention mechanisms (Bahdanau et al., 2014) have emerged as effective components in various neural network architectures. Attention works by aggregating a set of tokens via a weighted sum, where the attention weights are calculated as a function of both the input encodings and the state of the decoder.

Because attention mechanisms allocate weight among the encoded tokens, these coefficients are sometimes thought of intuitively as indicating which tokens the model focuses on when making a particular prediction. Based on this loose intuition, attention weights are often claimed to explain a model’s predictions. For example, in another work, De-Arteaga et al. (2019) study gender bias in machine learning models for occupation classification. As machine learning is increasingly used in hiring processes for tasks including resume filtering, the potential for bias on the basis of gender raises the specter that automating this process could lead to social harms. De-Arteaga et al. (2019) use attention over gender-revealing tokens (e.g., ‘she’, ‘he’, etc.) to verify the biases in occupation classification models—stating that “the attention weights indicate which tokens are most predictive”. Similar claims about attention’s utility for interpreting models’ predictions are common in the literature (Li et al., 2016; Xu et al., 2015; Choi et al., 2016; Xie et al., 2017; Martins and Astudillo, 2016; Lai and Tan, 2019).

In this paper, we question whether attention scores necessarily indicate features that influence a model’s predictions. Through a series of exper-
iments on diverse classification and sequence-to-sequence tasks, we show that attention scores are surprisingly easy to manipulate. We design a simple training scheme whereby the resulting models appear to assign little attention to a specified set of impermissible tokens while continuing to rely upon those features for prediction. The ease with which attention can be manipulated without significantly affecting performance suggests that even if a vanilla model’s attention weights conferred some insight (still an open and ill-defined question), these insights would rely on knowing the objective on which models were trained.

Our results present troublesome implications for proposed uses of attention in the context of fairness, accountability, and transparency. For example, malicious practitioners asked to justify how their models work by pointing to attention weights could mislead regulators with this scheme. For instance, looking at manipulated attention-based explanation in Table 1, one might (incorrectly) assume that the model does not rely on the gender prefix. To quantitatively study the extent of such deception, we conduct studies where we ask human subjects if the biased occupation classification models (like the ones audited by Arteaga et al. (2019)) rely on gender related information. We find that our manipulation scheme is able to deceive human annotators into believing that manipulated models do not take gender into account, whereas the models are heavily biased against gender minorities (see §5.2).

Lastly, practitioners often overlook the fact that attention is typically not applied over words but over final layer representations, which themselves capture information from neighboring words. We investigate the mechanisms through which the manipulated models attain low attention values. We note that (i) recurrent connections allow information to flow easily to neighboring representations; (ii) for cases where the flow is restricted, models tend to increase the magnitude of representations corresponding to impermissible tokens to offset the low attention scores; and (iii) models additionally rely on several alternative mechanisms that vary across random seeds (see §5.3).

2 Related Work

Many recent papers examine whether attention is a valid explanation or not. Jain et al. (2019) identify alternate adversarial attention weights after the model is trained that nevertheless produce the same predictions, and hence claim that attention is not explanation. However, these attention weights are chosen from a large (infinite up to numerical precision) set of possible values and thus it is not surprising that multiple weights produce the same prediction. Moreover since the model does not actually produce these weights, they would never be relied on as explanations in the first place. Similarly, Serrano and Smith (2019) modify attention values of a trained model post-hoc by hard-setting the highest attention values to zero. They find that the number of attention values that must be zeroed out to alter the model’s prediction is often too large, and thus conclude that attention is not a suitable tool to for determining which elements should be attributed as responsible for an output. In contrast to these two papers, we manipulate the attention via the learning procedure, producing models whose actual weights might deceive an auditor.

In parallel work to ours, Wiegrefe and Pinter (2019) examine the conditions under which attention can be considered a plausible explanation. They design a similar experiment to ours where they train an adversarial model, whose attention distribution is maximally different from the one produced by the base model. Here we look at a related but different question of how attention can be manipulated away from a set of impermissible tokens. We show that in this setting, our training scheme leads to attention maps which are more deceptive, since people find them to be more believable explanations of the output (see §5.2). We also extend our analysis to sequence-to-sequence tasks, and a broader set of models, including BERT, as well as identify mechanisms by which the manipulated models continue to rely on the impermissible tokens despite assigning low attention to them.

Lastly, several papers deliberately train attention weights by introducing an additional source of supervision to improve predictive performance. In some of these papers, the supervision comes from known word alignments for machine translation (Liu et al., 2016; Chen et al., 2016), or by aligning human eye-gaze with model’s attention for sequence classification (Barrett et al., 2018).

3 Manipulating Attention

Let \( S = w_1, w_2, \ldots, w_n \) denote an input sequence of \( n \) words. We assume that for each task, we are given a pre-specified set of impermissible words
For cases where the impermissible set of tokens is unknown apriori, one can plausibly use the top few highly attended tokens as a proxy.

4 Experimental Setup

We study the manipulability of attention on four binary classification problems, and four sequence-to-sequence tasks. In each dataset, (in some, by design) a subset of input tokens are known a priori to be indispensable for achieving high accuracy.

4.1 Classification Tasks

Occupation classification We use the biographies collected by De-Arteaga et al. (2019) to study bias against gender-minorities in occupation classification models. We carve out a binary classification task of distinguishing between surgeons and (non-surgeon) physicians from the multi-class
occupation prediction setup. We chose this sub-
task because the biographies of the two profes-
sions use similar words, and a majority of sur-
geons (> 80%) in the dataset are male. We further
downsampale minority classes—female surgeons,
and male physicians—by a factor of ten, to en-
courage models to use gender related tokens. Our
models (described in detail later in § 4.2) attain
96.4% accuracy on the task, and are reduced to
93.8% when the gender pronouns in the biogra-
phies are anonymized. Thus, the models (trained
on unanonymized data) make use of gender indi-
cators to obtain a higher task performance. Con-
sequently, we consider gender indicators as imper-
missible tokens for this task.

Pronoun-based Gender Identification We
construct a toy dataset from Wikipedia comprised
of biographies, in which we automatically label
biographies with a gender (female or male) based
solely on the presence of gender pronouns. To do
so, we use a pre-specified list of gender pronouns.
Biographies containing no gender pronouns, or
pronouns spanning both classes are discarded.
The rationale behind creating this dataset is that
due to the manner in which the dataset was
created, attaining 100% classification accuracy
is trivial if the model uses information from the
pronouns. However, without the pronouns, it may
not be possible to achieve perfect accuracy. Our
models trained on the same data with pronouns
anonymized, achieve at best 72.6% accuracy.

Sentiment Analysis with Distractor Sentences
We use the binary version of Stanford Senti-
ment Treebank (SST) (Socher et al., 2013), com-
prised of 10,564 movie reviews. We append
one randomly-selected “distractor” sentence to
each review, from a set of opening sentences of
Wikipedia pages. Here, without relying upon the
tokens in the SST sentences, a model should not be
able to outperform random guessing.

Graduate School Reference Letters We obtain
a dataset of recommendation letters written for the
purpose of admission to graduate programs. The
task is to predict whether the student, for whom
the letter was written, was accepted. The letters
include students’ ranks and percentile scores as
marked by their mentors, which admissions com-
nite members rely on. Indeed, we notice accu-

1Opening sentences tend to be declarative statements of
fact and typically are sentiment-neutral.
For the task of bigram flipping, the input lengths are restricted to be even. We use two sets of 100K unseen random sequences from the same distribution as the validation and test set.

**Machine Translation (English to German)** Besides synthetic tasks, we also evaluate on English to German translation. We use the Multi30K dataset, comprising of image descriptions (Elliott et al., 2016). Since the gold target to source word-level alignment is unavailable, we rely on the Fast Align toolkit (Dyer et al., 2013) to align target words to their source counterparts. We use these aligned words as impermissible tokens.

For all sequence-to-sequence tasks, we use an encoder-decoder architecture. Our encoder is a bidirectional GRU, and our decoder is a unidirectional GRU, with dot-product attention over source tokens, computed at each decoding timestep. We also run ablation studies with (i) no attention, i.e. just using the last (or the first) hidden state of the encoder; and (ii) uniform attention, i.e. all the source tokens are uniformly weighted.

**5 Results and Discussion**

In this section we examine how lowering attention affects task performance (§5.1). We then present experiments with human participants to quantify the deception with manipulated attention (§5.2). Lastly, we identify alternate workarounds through which models preserve task performance (§5.3).

**5.1 Attention mass and task performance**

For the classification tasks, we experiment with the loss coefficient \( \lambda \in \{0, 0.1, 1\} \). In each experiment, we measure the (i) attention mass: the sum of attention values over the set of impermissible tokens averaged over all the examples, and (ii) test accuracy. During the course of training (i.e. after each epoch), we arrive at different models from which we choose the one whose performance is within 2% of the original accuracy and provides the greatest reduction in attention mass on impermissible tokens. This is done using the development set, and the results on the test set from the chosen model are presented in Table 3. Across most tasks, and models, we find that our manipulation scheme severely reduces the attention mass on

---

2 These tasks have been previously used in the literature to assess the ability of RNNs to learn long-range reorderings and substitutions (Grefenstette et al., 2015).

---

3 Implementation details: the encoder and decoder token embedding size is 256, the encoder and decoder hidden dimension size is 512, and the teacher forcing ratio is 0.5. We use top-1 greedy strategy to decode the output sequence.

4 All data and code will be released on publication.
impermissible tokens compared to models without any manipulation (i.e. when $\lambda = 0$). This reduction comes at a minor, or no, decrease in task accuracy. Note that the models can not achieve performance similar to the original model (as they do), unless they rely on the set of impermissible tokens. This can be seen from the gap between models that do not use impermissible tokens ($I \times$) from ones that do ($I \checkmark$).

The only outlier to our findings is the SST+Wiki sentiment analysis task, where we observe that the manipulated Embedding and BiLSTM models reduce the attention mass but also lose accuracy. We speculate that these models are under-parameterized and thus jointly reducing attention mass and retaining original accuracy is harder. The more expressive BERT obtains an accuracy of over 90% while reducing the maximum attention mass over the movie review from 96.2% to $10^{-3}$%.

For **sequence-to-sequence tasks**, from Table 4, we observe that our manipulation scheme can similarly reduce attention mass over impermissible alignments while preserving original performance. To measure performance, we use token-by-token accuracy for synthetic tasks, and BLEU score for English to German MT. We also notice that the models with manipulated attention (i.e. deliberately misaligned) *outperform models with none or uniform attention*. This suggests that attention mechanisms add value to the learning process in sequence-to-sequence tasks which goes beyond their usual interpretation as alignments.

### 5.2 Human Study

We present three human subjects a series of inputs and outputs from the BiLSTM models, trained to predict occupation (physician or surgeon) given
a short biography.\textsuperscript{5} We highlight the input tokens as per the attention scores from three different schemes: (i) original dot-product attention, (ii) adversarial attention from Wiegreffe and Pinter (2019), and, (iii) our proposed attention manipulation strategy. We ask human annotators (Q1): Do you think that this prediction was influenced by the gender of the individual? Each participant answers either “yes” or “no” for a set of 50 examples from each of the three attention schemes.\textsuperscript{6} After looking at 50 examples from a given attention scheme, we inquire about trustworthiness of the attention scores (Q2): Do you believe the highlighted tokens capture the factors that drive the models’ prediction? They answer the question on a scale of 1 to 4, where 1 denotes that the highlighted tokens do not determine the models’ prediction, whereas 4 implies they significantly determine the models’ prediction. We deliberately ask participants once (towards the end) about the trustworthiness of attention-based explanations, in contrast to polling after each example, as it requires multiple examples to assess whether the explanations capture factors that are predictive.

Results We find that for the original dot-product attention, annotators labeled 66\% of predictions to be influenced by gender. Whereas for the other two attention schemes, none of the predictions were marked to be influenced by gender (see Table 5). This is despite all three models achieving roughly the same high accuracy (96\%) which relies on gender information. This demonstrates the efficacy of our manipulation scheme—predictions from models biased against gender minorities are perceived (by human participants) as not being influenced by gender. Further, our manipulated explanations receive a trustworthiness score of 2.67 (out of 4), only slightly lower than the score for the original explanations, and significantly better than the adversarial attention. We found that the KL divergence term in training adversarial attention (Eq. 1) encourages all the attention mass to concentrate on a single uninformative token for most examples, and hence was deemed as less trustworthy by the annotators (see Table 5, more examples in appendix). By contrast, our manipulation scheme only reduces attention mass over problematic tokens, and retains attention over non-problematic but predictive ones (e.g. “medicine”) making it more believable. We assess agreement among annotators, and calculate the Fleiss’ Kappa to be 0.97, suggesting almost perfect agreement.

5.3 Alternative Workarounds

We identify two mechanisms by which the models cheat, obtaining low attention values while remaining accurate.

Models with recurrent encoders can simply pass information across tokens through recurrent connections, prior to the application of attention. To measure this effect, we hard-set the attention values corresponding to impermissible words to zero after the manipulated model is trained, thus clipping their direct contributions for inference. For gender classification using the BiLSTM model, we are still able to predict over 99\% of instances correctly, thus confirming a large degree of information flow to neighboring representations.\textsuperscript{7} In contrast, the Embedding model (which has no means to pass the information pre-attention) attains only about 50\% test accuracy after zeroing the attention values for gender pronouns. We see similar evidence of passing around information in sequence-to-sequence models, where certain manipulated attention maps are off by one or two positions from the gold alignments (see Figure 2).

Models restricted from passing information prior to the attention mechanism tend to increase the magnitude of the representations corresponding to impermissible words to compensate for the

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
Attention & Example & Q1 & Q2 \\
\hline
Original & Ms. \textsuperscript{x} practices medicine and specializes in urological surgery & 66\% (yes) & 3.00 \\
\hline
Adversarial & Ms. \textsuperscript{x} practices medicine and specializes in urological surgery & 0\% (yes) & 1.00 \\
(Wiegreffe and Pinter, 2019) & & & \\
\hline
Ours & Ms. \textsuperscript{x} practices medicine and specializes in urological surgery & 0\% (yes) & 2.67 \\
\hline
\end{tabular}
\caption{Results to questions posed to human participants. Q1: Do you think that this prediction was influenced by the gender of the individual? Q2: Do you believe the highlighted tokens capture the factors that drive the models prediction? See § 5.2 for discussion.}
\end{table}

\textsuperscript{5}The participating subjects are graduate students, proficient in English, and unaware of our work.

\textsuperscript{6}We shuffled the order of sets among the three participants to prevent any ordering bias. Full details of the instructions presented to the annotators are in the appendix.

\textsuperscript{7} A recent study (Brunner et al., 2019) similarly observes a high degree of “mixing” of information across layers in Transformer models.
Figure 2: For three sequence-to-sequence tasks, we plot the original attention map on the left, followed by the attention plots of two manipulated models. The only difference between the manipulated models for each task is the (random) initialization seed. Different manipulated models resort to different alternative mechanisms.

Figure 3: For gender identification task, the norms of embedding vectors corresponding to impermissible tokens increase considerably in Embedding+Attention model to offset the low attention values. This is not the case for BiLSTM+Attention model as it can pass information due to recurrent connections.

We also notice that differently initialized models attain different alternative mechanisms. In Figure 2, we present attention maps from the original model, alongside two manipulated models initialized with different seeds. In some cases, the attention map is off by one or two positions from the gold alignments. In other cases, all the attention is confined to the first hidden state. In such cases, manipulated models are similar to a no-attention model, yet they offer better performance. In preliminary experiments, we found a few such models that outperform the no-attention baseline, even when the attention is turned off during inference. This suggests that attention offers benefits during training, even if it is not used during inference.

6 Conclusion

Amidst practices that perceive attention scores to be an indication of what the model focuses on, we show that attention scores are easily manipulable. Our simple training scheme produces models with significantly reduced attention mass over tokens known a priori to be useful for prediction, while continuing to use them. Our results raise concerns about the potential use of attention as a tool to audit algorithms, as malicious actors could employ similar techniques to mislead regulators.
References

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

Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. 2018. Sequence classification with human attention. In Proceedings of the 22nd Conference on Computational Natural Language Learning, pages 302–312.

Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, and Roger Wattenhofer. 2019. On the validity of self-attention as explanation in transformer models. arXiv preprint arXiv:1908.04211.

Wenhu Chen, Evgeny Matusov, Shahram Khadivi, and Jan-Thorsten Peter. 2016. Guided alignment training for topic-aware neural machine translation. arXiv preprint arXiv:1607.01628.

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.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. arXiv preprint arXiv:1901.09451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. North American Chapter of the Association for Computational Linguistics (NAACL).

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. North American Chapter of the Association for Computational Linguistics (NAACL).

Desmond Elliott, Stella Frank, Khalil Sima’an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. arXiv preprint arXiv:1605.00459.

Andrea Galassi, Marco Lippi, and Paolo Torroni. 2019. Attention, please! A critical review of neural attention models in natural language processing. arXiv preprint arXiv:1902.02181.

Alex Graves and Jürgen Schmidhuber. 2005. Frame-wise phoneme classification with bidirectional lstm and other neural network architectures. Neural Networks, 18(5-6):602–610.

Edward Grefenstette, Karl Moritz Hermann, Mustafa Suleyman, and Phil Blunsom. 2015. Learning to transduce with unbounded memory. In Advances in neural information processing systems, pages 1828–1836.

Sarthak Jain, Ramin Mohammadi, and Byron C Wallace. 2019. Attention is not explanation. North American Chapter of the Association for Computational Linguistics (NAACL).

Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pages 29–38. ACM.

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

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Neural machine translation with supervised attention. arXiv preprint arXiv:1609.04186.

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.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? 57th annual meeting of the Association for Computational Linguistics (ACL).

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.

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.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. Proceedings of the 2019 conference on Empirical Methods in Natural Language Processing, EMNLP.

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

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. arXiv preprint arXiv:1502.03044.