Is Sparse Attention more Interpretable?

Meister, Clara; Lazov, Stefan; Augenstein, Isabelle; Cotterell, Ryan

Published in:
Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)

DOI:
10.18653/v1/2021.acl-short.17

Publication date:
2021

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Meister, C., Lazov, S., Augenstein, I., & Cotterell, R. (2021). Is Sparse Attention more Interpretable? In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 122-129). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.17
Is Sparse Attention more Interpretable?

Clara Meister† Stefan Lazov* Isabelle Augenstein‡ Ryan Cotterell*†
†ETH Zürich  *University of Cambridge  ‡University of Copenhagen
meistecl@inf.ethz.ch, stefan.lazov@cantab.net, augenstein@di.ku.dk, ryan.cotterell@inf.ethz.ch

Abstract
Sparse attention has been claimed to increase model interpretability under the assumption that it highlights influential inputs. Yet the attention distribution is typically over representations internal to the model rather than the inputs themselves, suggesting this assumption may not have merit. We build on the recent work exploring the interpretability of attention; we design a set of experiments to help us understand how sparsity affects our ability to use attention as an explainability tool. On three text classification tasks, we verify that only a weak relationship between inputs and co-indexed intermediate representations exists—under sparse attention and otherwise. Further, we do not find any plausible mappings from sparse attention distributions to a sparse set of influential inputs through other avenues. Rather, we observe in this setting that inducing sparsity may make it less plausible that attention can be used as a tool for understanding model behavior.

1 Introduction
Interpretability research in natural language processing (NLP) is becoming increasingly important as complex models are applied to more and more downstream decision making tasks. In light of this, many researchers have turned to the attention mechanism, which has not only led to impressive performance improvements in neural models, but has also been claimed to offer insights into how models make decisions. Specifically, a number of works imply or directly state that one may inspect the attention distribution to determine the amount of influence each input token has in a model’s decision-making process (Xie et al., 2017; Mullenbach et al., 2018; Niculae et al., 2018; inter alia).

Many lines of work have gone on to exploit this assumption when building their own “interpretable” models or analysis tools (Yang et al., 2016; Tu et al., 2016; De-Arteaga et al., 2019); one subset has even tried to make models with attention more interpretable by inducing sparsity—a common attribute of interpretable models (Lipton, 2018; Rudin, 2019)—in attention weights, with the motivation that this allows model decisions to be mapped to a limited number of items (Martins and Astudillo, 2016; Malaviya et al., 2018; Zhang et al., 2019). Yet, there lacks concrete reasoning or evidence that sparse attention weights leads to more interpretable models: customarily, attention is not directly over the model’s inputs, but rather over some representation internal to the model, e.g. the hidden states of a recurrent network or contextual embeddings of a Transformer (see Fig. 1). Importantly, these internal representations do not solely encode information from the input token they are co-indexed with (Salehinejad et al., 2017; Brunner et al., 2020), but rather from a range of inputs. This presents the question: if internal representations themselves may not be interpretable, can we actually deduce anything from “interpretable” attention weights?

We build on the recent line of work challenging the validity of attention-as-explanation methods (Jain and Wallace, 2019; Serrano and Smith, 2019; Grimsley et al., 2020, inter alia) and specifically examine how sparsity affects their observations. To this end, we introduce a novel entropy-based metric to measure the dispersion of inputs’ influence, rather than just their magnitudes. Through experiments on three text classification tasks, utilizing both LSTM and Transformer-based models, we observe how sparse attention affects the results of Jain and Wallace (2019) and Wiegreffe and Pinter (2019), additionally exploring whether it allows us to identify a core set of inputs that are important to models’ decisions. We find we are unable to identify such a set when using sparse attention; rather, it appears that encouraging sparsity may simultaneously encourage a higher degree of
contextualization in intermediate representations. We further observe a decrease in the correlation between the attention distribution and input feature importance measures, which exacerbates issues found by prior works. The primary conclusion of our work is that we should not believe sparse attention enhances model interpretability until we have concrete reasons to believe so; in this preliminary analysis, we do not find any such evidence.

### 2 Attention-based Neural Networks

We consider inputs $x = x_1 \cdots x_n \in V^n$ of length $n$ where the tokens from taken from an alphabet $V$. We denote the embedding of $x$, e.g., its one hot encoding or (more commonly) a linear transformation of its one-hot encoding with an embedding matrix $E \in \mathbb{R}^{d \times |V|}$, as $X(e) \in \mathbb{R}^{d \times n}$. Our embedded input $X(e)$ is then fed to an encoder, which produces $n$ intermediate representations $I = [h_1; \ldots; h_n] \in \mathbb{R}^{m \times n}$, where $h_i \in \mathbb{R}^m$ and $m$ is a hyperparameter of the encoder. This structure is quite architecture dependent.

An alignment function $A(\cdot, \cdot)$ maps a query $q$ and a key $K$ to weights $a(t)$ for a decoding time step $t$; we subsequently drop $t$ for simplicity. In colloquial terms, $A$ chooses which values of $K$ should receive the most attention based on $q$, which is then represented in the vector $a(t) \in \mathbb{R}^n$ for $t$. For the NLP tasks we consider, we have $K = I = [h_1; \ldots; h_n]$, the encoder outputs. A query $q$ may be, e.g., a representation of the question in question answering.

The weights $a$ are projected to sum to 1, which results in the attention distribution $\alpha$. Mathematically, this is done via a projection onto the probability simplex using a function $\phi$, e.g., softmax or sparsemax. We then compute the context vector as $c = \sum_{i=1}^n \alpha_i h_i$. This context vector is fed to a decoder, whose structure is again architecture dependent, which generates a (possibly unnormalized) probability distribution over the set of labels $\mathcal{Y}$, where $\mathcal{Y}$ is defined by the task.

**Attention.** We experiment with two methods of constructing an attention distribution: (1) additive attention, proposed by Bahdanau et al. (2015): $A(K, q)_i = v^\top \tanh(W_1 K_i + W_2 q)$, and (2) the scaled dot product alignment function, as in the Transformer network: $A(K, q) = K^\top q / \sqrt{m}$ where $v \in \mathbb{R}^d$ and $W_1, W_2 \in \mathbb{R}^{d \times n}$ are weight matrices. Note that the original (without attention) neural encoder–decoder architecture, as in Sutskever et al. (2014), can be recovered with alignment function $A(\cdot, \cdot) = [0, \ldots, 0, 1]$, i.e., only the last of the $n$ intermediate representations is given to the decoder.

**Projection Functions.** A projection function $\phi$ takes the output of the alignment function and maps it to a valid probability distribution: $\phi : \mathbb{R}^n \rightarrow \Delta^{n-1}$. The standard projection function is softmax:

$$
\phi_{\text{soft}}(z) = \frac{\exp(z)}{\sum_{i \in [n]} \exp(z_i)} \quad (1)
$$

However, softmax leads to non-sparse solutions as an entry $\phi_{\text{soft}}(z)_i$ can only be 0 if $x_i = -\infty$. Alternatively, Martins and Astudillo (2016) introduce **sparsemax**, which can output sparse distributions:

$$
\phi_{\text{sparse}}(z) = \arg\min_{p \in \Delta^{n-1}} \|p - z\|_2^2 \quad (2)
$$

In words, sparsemax directly maps $z$ onto the probability simplex, which often leads to solutions on the boundary, i.e., where at least one entry of $p$ is 0. One shortcoming of sparsemax is the lack of control over the degree of sparsity. **Sparsegen** (Laha et al., 2018) fills this gap:

$$
\phi_{\text{sparsegen}}(z) = \arg\min_{p \in \Delta^{n-1}} \|p - g(z)\|_2^2 - \lambda \|p\|_2^2 \quad (3)
$$

where the degree of sparsity can be tuned via the hyperparameter $\lambda \in (-\infty, 1)$; a larger $\lambda$ encourages more sparsity in the minimizing solution.

### 3 Model Interpretability

Model interpretability and explainability have been framed in different ways (Gehrmann et al., 2019)—as model understanding tasks, where (spurious) features learned by a model are identified, or as decision understanding tasks, where explanations for particular instances are produced. We consider the latter in this paper. Such tasks can be framed as generative, where models generate free text explanations (Camburu et al., 2018; Kotonya and Toni, 2020; Atanasova et al., 2020b), or as post-hoc interpretability methods, where salient portions of the input are highlighted (Lipton, 2018; DeYoung et al., 2020; Atanasova et al., 2020a).

As there does not exist a clearly superior choice for framing decision understanding for NLP tasks (Miller, 2019; Carton et al., 2020; Jacovi and...
Goldberg, 2021), we follow a substantial body of prior work in considering the post-hoc definition of interpretability based on local methods proposed by Lipton (2018). This definition is naturally operationalized through feature importance metrics and meta-models (Jacovi and Goldberg, 2020). Further, we acknowledge the specific requirement that an interpretable model obeys some set of structural constraints of the domain in which it is used, such as monotonicity or physical constraints (Rudin, 2019). For NLP tasks such as sentiment analysis or topic classification, such constraints may logically include the utilization of only a few key words in the input when making a decision, in which case, knowing the magnitude of the influence each input token has on a model’s prediction through, e.g., feature importance metrics, may suffice to verify the model obeys such constraints. While this collective definition is limited (Doshi-Velez and Kim, 2017; Guidotti et al., 2018; Rudin, 2019), we posit that if attention cannot provide model interpretability at this level, then it would likewise not be able to under more rigorous constraints.

3.1 Measures of Feature Importance

Gradient-Based Methods. Gradient-based measures of feature importance (F1; Baehrens et al., 2010; Simonyan et al., 2014; Poermer et al., 2018) use the gradient of a function’s output w.r.t. a feature to measure the importance of that feature. In the case of an attentional neural network for binary classification \( f(\cdot) \), we can take the gradient of \( f \) w.r.t. the variable \( x \) and evaluate at a point \( x = x' \) to gain a sense of how much influence each \( x_i \) had on the outcome \( \hat{y} = f(x') \). These measures are not restricted to the relationship between inputs \( x_i \) and the outcome \( f(x) \); they can also be adapted to measure for effects from and to intermediate representations \( h_p \). Formally, our measures are as follows:

\[
g_{\hat{y}}(x_i) = \left\| \frac{\partial f}{\partial x_i} \right\|_2 \sum_{k=1}^{n} \left\| \frac{\partial f}{\partial x_k} \right\|_2 \tag{4}
\]

\[
g_{h_p}(x_i) = \left\| \frac{\partial \|h_p\|_2}{\partial x_i} \right\|_2 \sum_{k=1}^{n} \left\| \frac{\partial \|h_p\|_2}{\partial x_k} \right\|_2 \tag{5}
\]

where \( g_{\hat{y}}(x_i) \in [0, 1] \) and \( g_{h_p}(h_p) \in [0, 1] \) represents the gradient-based FI of token \( x_i \) on \( \hat{y} \) and intermediate representation \( h_p \), respectively.

Gradient-based methods are often used in explainability techniques, as they have exhibited higher correlation with human judgement than others (Atanasova et al., 2020a). Note that we take gradients w.r.t. the embedding of token \( x_i \) and that in the latter metric, we measure the influence of \( x_i \) on the magnitude of \( h_p \)—a decision we discuss in App. A.

Leave-One-Out (LOO)-based Methods. As a secondary FI metric, we observe how model predictions change when a specific input token is removed. For token \( x_i \), this can be calculated as:

\[
D_{\hat{y}}(x_i) = \frac{|\hat{y} - \hat{y}_{-i}|}{\sum_{k=1}^{n} |\hat{y} - \hat{y}_{-k}|} \tag{6}
\]

where \( \hat{y}_{-i} \) is the prediction of a model with input \( x_i \) removed. The formula can also be used for intermediate representations; we denote this as \( D_{\hat{y}}(h_i) \).

4 Experiments

Setup. We run experiments across several model architectures, attention mechanisms, and datasets, in order to understand the effects of induced attentional sparsity on model interpretability. We use three binary classification datasets: IMDB and SST (sentiment analysis) and 20News (topic classification). We use the dataset versions provided by Jain and Wallace (2019), exactly following their pre-processing steps. We show a subset of representative results here, with additional results in App. C. Further details, including model architecture descriptions, dataset statistics, and baselines accuracies may be found in App. B.

Inputs and Intermediate Representations are not Interchangeable. We first explore how strongly-related inputs are to their co-indexed intermediate representations. A strong relationship on its own may validate the use of sparse attention, as the ability to identify a subset of influential intermediate representations would then directly translate to a set of influential inputs. Previous works show that the “contribution” of a token \( x_i \) to its intermediate representation \( h_i \) is often quite low for various model architectures (Salehinejad et al., 2017; Ming et al., 2017; Brunner et al., 2020; Tutek and Snajder, 2020). In the context of attention, we find this property to be evinced by the adversarial experiments of Wiegreffe and Pinter (2019) (§4) and Jain and Wallace (2019) (§4), which we verify in App. C. They construct adversarial attention distributions by optimizing for divergence from a baseline
Figure 1: Correlation between the attention distribution and gradient-based FI measures. We see a notably stronger correlation between attention and FI of intermediate representation than of inputs across all models.

| Dataset | IMdb | 20-News | SST |
|---------|------|---------|-----|
| BILSTM (Softmax) | 0.71 ± 0.09 | 0.75 ± 0.12 | 0.93 ± 0.05 |
| BILSTM (Sparsemax) | 0.72 ± 0.10 | 0.68 ± 0.12 | 0.91 ± 0.07 |
| Transformer (Softmax) | 0.76 ± 0.08 | 0.48 ± 0.06 | 0.73 ± 0.09 |
| Transformer (Sparsemax) | 0.72 ± 0.09 | 0.40 ± 0.06 | 0.63 ± 0.08 |

Table 1: Mean entropy of gradient-based FI of input to intermediate representations. Green numbers are std. deviations. Projection functions are parenthesized.

As further verification of high degrees of contextualization in attentional models, we report a novel quantification, offering insights into whether individual intermediate representations can be linked primarily to any single input—i.e., perhaps not the co-indexed input; we explore whether a subset can still be identified through FI metrics. In the case where the normalized FI distribution highlights only a few key items, the distribution will, by definition, have low entropy. Thus, we explore whether sparse attention leads to lower entropy input FI distributions in comparison to standard attention. We find no such trend; Fig. 2 shows that across all models and settings, the entropy of the FI distribution is quite high. Further, we see a consistent negative correlation between this entropy and the sparsity parameter of the sparsegen projection (Table 2), implying that entropy of feature importance increases as we raise the degree of sparsity in  $\lambda$.

Table 2: Correlation between sparsegen parameter $\lambda$ and entropy of gradient-based input FI $\hat{H}(g_i(x))$.

| Dataset | IMdb | 20-News | SST |
|---------|------|---------|-----|
| BILSTM (tanh) | -0.935 | -0.675 | -0.866 |
| Transformer (dot) | -0.830 | -0.409 | -0.810 |

Sparse Attention $\neq$ Sparse Input Feature Importance. Our prior results demonstrated that—even when using sparse attention—we cannot identify a subset of influential inputs directly through intermediate representations; we explore whether a subset can still be identified through FI metrics. In the case where the normalized FI distribution highlights only a few key items, the distribution will, by definition, have low entropy. Thus, we explore whether sparse attention leads to lower entropy input FI distributions in comparison to standard attention. We find no such trend; Fig. 2 shows that across all models and settings, the entropy of the FI distribution is quite high. Further, we see a consistent negative correlation between this entropy and the sparsity parameter of the sparsegen projection (Table 2), implying that entropy of feature importance increases as we raise the degree of sparsity in $\alpha$.

Correlation between Attention and Feature Importance. Finally, we follow the experimental setup of Jain and Wallace (2019), who postulate...
Figure 3: Correlation between the attention distribution and input FI measures as a function of the sparsity penalty $\lambda$ used in the projection function $\phi_{\text{spars}}$. The $x$-axis is log-scaled for $\lambda < 0$ since $\lambda \in (-\infty, 1)$. Results are from the IMDb dataset.

that if the attention distribution indicates which inputs influence model behavior, then one may reasonably expect attention to correlate with FI measures of the input. While they find only a weak correlation, we explore how inducing sparsity in the attention distribution affects this result. Surprisingly, Fig. 3 shows a downward trend in this correlation as the sparsity parameter $\lambda$ of the sparsgen projection function is increased. As argued by Wiegrefe and Pinter (2019), a lack of this correlation does not indicate attention cannot be used as explanation; FI measures are not ground-truth indicators of critical inputs. However, the inverse relationship between input FI and attention is rather surprising. If anything, we may surmise sparsity in $\alpha$ leads to less faithful explanations from $\alpha$. From these results, we posit that promoting sparsity in attention distribution may simply lead to the dispersion of information to different intermediate representations, a behavior similar to that seen when constraining attention for divergence from another distribution, i.e., in the adversarial experiments of Wiegrefe and Pinter (2019) compared to those of Jain and Wallace (2019).

5 Related Work

The use of attention as an indication of inputs’ influence on model decisions may at first seem natural; yet a large body of work has recently challenged this practice. Perhaps the first to do so was Jain and Wallace (2019), which revealed both a lack of correlation between the attention distribution and well established feature importance metrics and of unique optimal attention weights. Subsequently, other studies arrived at similar results: Grimsley et al. (2020) found evidence that causal explanations are not attainable from attention layers over text data; Jacovi and Goldberg (2020) explored the faithfulness of attention heatmaps; Fruthi et al. (2020) showed that attention masks can be trained to give deceptive explanations. We view this work as another such study, exploring attention’s innate interpretability on a different axis.

Further, this work fits into the context of a larger body of interpretability research in NLP, which has challenged the informal use of terms such as faithfulness, plausibility, and explainability (Lipton, 2018; Arrieta et al., 2020; Jacovi and Goldberg, 2021, inter alia) and tried to quantify the reliability of current definitions (Atanasova et al., 2020a). While we consider their findings in our experimental design—e.g., in our choice of feature importance metrics—we recognize that further experiments would be needed to address all of their concerns; for example, this work could be extended by using the benchmark created by DeYoung et al. (2020) as an additional metric of interpretability.

6 Conclusion

Prior work has cited interpretability as a driving factor for promoting sparsity in attention distributions. We explore how induced sparsity affects the ability to use attention as a tool for explaining model decisions. In our experiments on text classification tasks, we see that while sparse attention distributions may allow us to pinpoint influential intermediate representations, we are unable to find any plausible mapping from sparse attention to a small, critical set of influential inputs. Rather, we find evidence that inducing sparsity may make it even less plausible to use the attention distribution to interpret model behavior. We conclude that we need further reason to believe sparse attention increases model interpretability as our results do not support such claims.

Acknowledgements

We thank the anonymous reviewers for their insightful feedback on the manuscript. Isabelle Augenstein’s research is partially funded by a DFF Sapere Aude research leader grant.

---

2We use Kendall’s $\tau$-b correlation (Knight, 1966).

3Serrano and Smith (2019) contemporaneously found similar results.
Ethical Considerations

Machine learning models are being deployed in an increasing number of sensitive situations. In these settings, it is critical that models are interpretable, so that we can avoid e.g., inadvertent racial or gender bias. Giving a false sense of interpretability can allow models with undesirable (i.e., unethical or unstable) behavior to fly under the radar. We view this work as another critique of interpretability claims and hope our results will encourage the more careful consideration of interpretability assumptions when using machine learning models in practice.

References

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion, 58.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lima, and Isabelle Augenstein. 2020a. A diagnostic study of explainability techniques for text classification. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3256–3274, Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lima, and Isabelle Augenstein. 2020b. Generating fact checking explanations. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to explain individual classification decisions. Journal of Machine Learning Research, 11:1803–1831.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In 3rd International Conference on Learning Representations.

Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in Transformers. In 8th International Conference on Learning Representations.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In Advances in Neural Information Processing Systems 31, pages 9539–9549. Curran Associates, Inc.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9294–9307, Association for Computational Linguistics.

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. In Proceedings of the Conference on Fairness, Accountability, and Transparency, page 120–128, Association for Computing Machinery.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4443–4458, Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. CoRR, abs/1702.08608.

Sebastian Gehrmann, Hendrik Strobelt, Robert Krueger, Hanspeter Pfister, and Alexander M Rush. 2019. Visual interaction with deep learning models through collaborative semantic inference. IEEE Transactions on Visualization and Computer Graphics, 26(1):884–894.

Christopher Grimsley, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. ACM Computing Surveys, 51(5).

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4198–4205, Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2021. Aligning Faithful Interpretations with their Social Attribution. Transactions of the Association for Computational Linguistics, 9:294–310.

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 Lan-
Martin Tutek and Jan Snajder. 2020. Staying true to your word: (how) can attention become explanation? In Proceedings of the 5th Workshop on Representation Learning for NLP, pages 131–142, Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 11–20, Association for Computational Linguistics.

Qizhe Xie, Xuezhe Ma, Zihang Dai, and Eduard Hovy. 2017. An interpretable knowledge transfer model for knowledge base completion. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 950–962, Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1480–1489, Association for Computational Linguistics.

J. Zhang, Y. Zhao, H. Li, and C. Zong. 2019. Attention with sparsity regularization for neural machine translation and summarization. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(3):507–518.