UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Understanding Multi-Head Attention in Abstractive Summarization

Baan, J.; ter Hoeve, M.; van der Wees, M.; Schuth, A.; de Rijke, M.

DOI
10.48550/arXiv.1911.03898

Publication date
2019

Document Version
Submitted manuscript

Link to publication

Citation for published version (APA):
Baan, J., ter Hoeve, M., van der Wees, M., Schuth, A., & de Rijke, M. (2019). *Understanding Multi-Head Attention in Abstractive Summarization*. (v1 ed.) ArXiv. https://doi.org/10.48550/arXiv.1911.03898

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:13 Oct 2026

Understanding Multi-Head Attention in Abstractive Summarization

Joris Baan[1] **Maartje ter Hoeve**[2] **Marlies van der Wees**[1]
Anne Schuth[1] **Maarten de Rijke**[2]
[1]DPG Media, Amsterdam [2]University of Amsterdam, Amsterdam
{joris.baan,marlies.van.der.wees,anne.schuth}@dpgmedia.nl
{m.a.terhoeve,derijke}@uva.nl

arXiv:1911.03898v1 [cs.CL] 10 Nov 2019

Abstract

Attention mechanisms in deep learning architectures have often been used as a means of transparency and, as such, to shed light on the inner workings of the architectures. Recently, there has been a growing interest in whether or not this assumption is correct. In this paper we investigate the interpretability of multi-head attention in abstractive summarization, a sequence-to-sequence task for which attention does not have an intuitive alignment role, such as in machine translation. We first introduce three metrics to gain insight in the focus of attention heads and observe that these heads specialize towards relative positions, specific part-of-speech tags, and named entities. However, we also find that ablating and pruning these heads does not lead to a significant drop in performance, indicating redundancy. By replacing the softmax activation functions with sparsemax activation functions, we find that attention heads behave seemingly more transparent: we can ablate fewer heads and heads score higher on our interpretability metrics. However, if we apply pruning to the sparsemax model we find that we can prune even more heads, raising the question whether enforced sparsity actually improves transparency. Finally, we find that relative positions heads seem integral to summarization performance and persistently remain after pruning.

1 Introduction

As learning algorithms become more powerful, their role in important decision making grows. At the same time the complexity of these learning algorithms increases (Schmidhuber, 2015). This has given rise to a strong demand for more transparency in deep learning models, both from the general public (Voigt and Von dem Bussche, 2017) and the research community (e.g., Doshi-Velez and Kim, 2017; Miller, 2018). Attention (Bahdanau et al., 2014; Luong et al., 2015) has gained popularity as a means of obtaining insight in the inner workings of deep neural networks (e.g., Lei, 2017; Choi et al., 2016; Gilpin et al., 2018; Ghaeini et al., 2018). Often examples of the attention heat map are provided to point out what attention focuses on. However, these examples are typically cherry-picked and leave it unclear to what extent attention can be used for transparency. In fact, a growing body of research has shown that one should use caution when using attention as a means for transparency. The majority of this work has focused on Machine Translation (e.g., Vashishth et al., 2019) and classification (e.g., Jain and Wallace, 2019), however, interpretability of attention for other tasks has not been researched as thoroughly.

We argue that in order to get a full understanding of the interpretability of attention, we should broaden our focus to other areas. Therefore, in this work we focus on abstractive summarization, a task that is particularly interesting for analyzing transparency since the correspondence between input and output is less clear than in machine translation. Yet due to the sequence-to-sequence nature of the task the benefit of attention is more apparent than for language classification tasks. We specifically focus on the state-of-the-art transformer architectures (Vaswani et al., 2017) that are commonly used for this task. By doing so, we contribute as follows:

- (**C1**) We introduce new metrics that can be used to evaluate transparency in abstractive summarization.
- (**C2**) We provide insights in what attention heads in state of the art transformer architectures focus on in abstractive summarization.
- (**C3**) We analyze two methods (inducing sparsity and pruning) for increasing multi-head attention interpretability applied to abstractive summarization.

2 Defining Transparency, Explainability Interpretability and Faithfulness

Before we analyze transparency, explainability, interpretability and faithfulness, we need clear definitions of each of these concepts. Doshi-Velez and Kim (2017) define interpretability in Machine Learning as "*the ability to explain or to present in understandable terms to a human*". Gilpin et al. (2018) define an explanation to be an answer to "why questions" and consider it a trade-off between interpretability and completeness. Interpretability here means *understandable to humans*, whereas completeness covers how well the explanation is *faithful* to the actual model mechanics. Intuitively, a *transparent* model is a model in which it is clear what is happening inside. However, simply providing all parameters along with the optimization procedure violates this intuition and appears to be cheating. A transparent model should thus also be interpretable to some degree. The exact difference between explainability and transparency remains illusive. We argue that transparency addresses the **what** question: what is happening within the model? In contrast, following Gilpin et al.'s definition, explainability addresses the **why** question: why is this output produced? Both transparency and explainability should be evaluated in terms of interpretability (how easily can we understand this explanation?) as well as faithfulness (how well does this explanation describe the system in an accurate way?).

3 Related Work

3.1 Abstractive summarization

The field of summarization can be divided in *extractive* and *abstractive* summarization. In this work we focus on the latter. The task of abstractive summarization is to construct summaries by generating new words and sentences, as opposed to directly extracting parts from the source text to add to the summary (which is extractive summarization). Deep learning has helped to advance abstractive summarization (e.g., Rush et al., 2015; Nallapati et al., 2016; See et al., 2017; Narayan et al., 2018). With the introduction of the transformer (Vaswani et al., 2017) and representation models like BERT (Devlin et al., 2018), performance on the abstractive summarization task has increased again (e.g., Gehrmann et al., 2018; Liu and Lapata, 2019). We follow the state of the art and focus on transformer architectures for abstractive summarization.

3.2 Attention as an interpretability metric

Recently there has been a lot of interest in whether or not attention can be used to interpret or explain a model's inner functionality. Some of this work argues it can (e.g., Vig and Belinkov, 2019; Clark et al., 2019; Correia et al., 2019), whereas other work argues it cannot, or one should at least be cautious (e.g., Jain and Wallace, 2019; Serrano and Smith, 2019). Vashishth et al. (2019) analyze the problem over a variety of NLP tasks and conclude that it depends on the task and the importance of attention for this task. For tasks where attention does not seem to play an important role (such as text classification), attention cannot be used as interpretability metric, whereas for other tasks where attention does play a major role (such as machine translation) it can. None of these previous works focus on abstractive summarization – a sequence-to-sequence task where attention is beneficial, yet expected to behave differently than the attention in machine translation as the correspondence between input and output is less straight forward. We close this gap in this work.

3.3 Ablation and pruning of attention heads

Michel et al. (2019) show that a large number of attention heads can be removed without a significant drop in model performance when applying the transforme to machine translation and BERT (Devlin et al., 2018) to natural language inference tasks. Voita et al. (2019) introduce a pruning method that we also use in our research, hence we describe it in more detail here. Voita et al. apply a strategy that allows the model to retrain itself. They augment multi-head attention (MHA) with gates and consider them head-specific model parameters in the closed $[0, 1]$ interval. The objective is to encourage the model to shut down heads by pushing their gates to exactly zero. Voita et al. use a stochastic relaxation of the L_0 norm as follows:

$$L_C(\boldsymbol{\phi}) = \sum_{i=1}^{h} (1 - Q_{\bar{s}_i}(g_i = 0 | \phi_i)) \quad (1)$$

$$= \sum_{i=1}^{h} sigmoid(\log \alpha_j - \beta \log \frac{\epsilon}{1+\epsilon}) \quad (2)$$

L_C approximates the number of non-zero gates using the probability of these gates being non-zero. During training, gates are individually sampled from a Hard Concrete distribution (Louizos et al., 2017), of which the distribution parameter $\log \alpha$ is learned. Gates are resampled for each batch. The coefficient λ controls the weight of the regularization penalty. During inference, fixed gate values are obtained through:

$$\hat{\boldsymbol{g}} = \min(\mathbf{1}, \max(\mathbf{0}, \\ sigmoid(\log \boldsymbol{\alpha})(1 + 2\epsilon) - \epsilon)). \quad (3)$$

Voita et al. find that the majority of the heads can be pruned with a minimal effect on overall translation performance (BLUE). This method has not been applied to the task of abstractive summarization.

3.4 Sparsity in attention

Sparsity has been used to improve the interpretability of single-head attention architectures (Malaviya et al., 2018; Deng et al., 2018; Niculae et al., 2018). Commonly, these methods are based on, or extend, a sparsemax transformation (Martins and Astudillo, 2016). Recently, Correia et al. (2019) apply an extension of sparsemax to multi-head attention architectures for NMT. Correia et al. (2019) replace the softmax in the attention heads with an α-entmax; the higher the value for α the sparser. They show for a number of metrics that the model becomes more interpretable.

4 Experimental Setup

We use the *CNN/Daily Mail* (Hermann et al., 2015; Nallapati et al., 2016) data set which consists of news articles: 287,226 training, 13,368 validation and 11,490 test pairs. Articles consist on average of 781 tokens and summaries of 3.75 sentences or 56 tokens. Following See et al. (2017) we truncate articles to 400 words. We recover the original capitalized articles to better identify part-of-speech (POS) and named entity (NE) tags, as current state-of-the-art taggers are sensitive to capitalization. To obtain the named entity and part-of-speech tags used for our analysis, we use out-of-the-box taggers by Akbik et al. (2018).[1]

[1] https://github.com/zalandoresearch/flair

	R-1 F1	R-2 F1	R-L F1
Model 1	38.76	17.13	36.00
Model 2	38.81	16.77	36.28

Table 1: ROUGE scores for two identical models with a different parameter initialization seed.

We adopt OpenNMT's implementation (Klein et al., 2017) of the CopyTransformer (Gehrmann et al., 2018) with its default hyper-paramameters. The encoder and decoder have four layers with eight heads per layer. We use two architecturally identical models. The first is an out-of-the-box model that has been pre-trained by Klein et al. (2017).[2] The second is an identical model with different parameter initialization to investigate whether stochasticity affects the way attention heads specialize. Table 1 shows the summarization performance of both models measured in ROUGE (Lin, 2004).

5 Analyzing Multi-head Attention

We start our investigation by visually inspecting heatmaps of attention distributions to gain an intuition of its behavior (see an example heatmap in Figure 1). We observe that some heads focus on locations, people or key words—all word types that seem important to summarization.

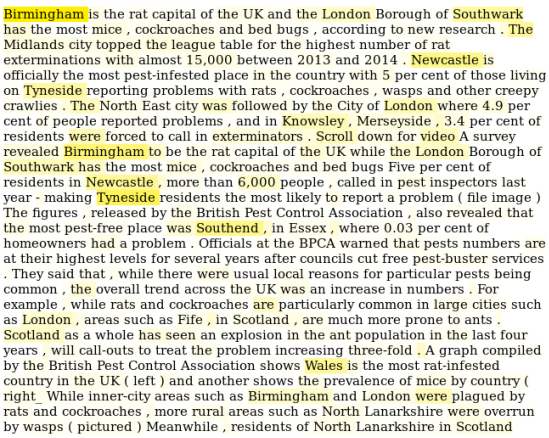


Figure 1: Attention heatmap for a decoder head that focuses on locations.

To quantitatively verify this observation, we design three metrics, measuring syntactic (§5.1), semantic (§5.2), and positional (§5.3) patterns, respectively. We apply these metrics to the attention distributions generated during summarization. We analyze heads from the encoder (self-attention) as

[2] http://opennmt.net/OpenNMT-py/Summarization.html

well as the decoder (cross-attention) using 1K randomly sampled news articles that we pre-tag with part-of-speech and named entity tags. We exclude decoder self-attention because (i) its attention spans increase step-wise, causing a quantitative analysis to be significantly more complex, and (ii) encoder self-attention and cross-attention are more commonly used for interpreting MHA (Raganato et al., 2018; Clark et al., 2019; Vaswani et al., 2017). We examine two identical models with a different random seed in Section 5.4 and the importance of individual heads in Section 5.5.

The work presented in this section expands upon previous exploratory work by Baan et al. (2019) on analyzing transformer heads in abstractive summarization. Two key additions are the visualizations in Figure 3 and the ablation study in Section 5.5.

5.1 Syntactic patterns

To quantify syntactic patterns, we compute the average KL-divergence between normalized POS tag histograms of articles and normalized attention-weighted POS tag histograms (POS-KL). A head focusing on specific POS tags will obtain a high POS-KL (see Figure 3a). Decoder heads seem more specialized towards syntax, reflected by more heads with a high POS-KLs. These heads also correspond to 'syntax' heads in our qualitative analysis. The peaks at the nouns and punctuation tags in Figure 2 provide insight into what exactly these heads specialize towards. However, the KL-divergences are relatively low (below 0.5) and we observe that there is still a considerable portion of probability mass on other syntactic categories. This means that heads do focus on syntax, but that it not generalize perfectly over 1000 articles.

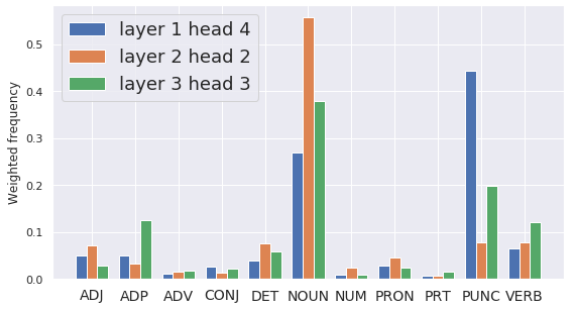


Figure 2: Attention distribution over POS tags for the top three specialized syntactic decoder heads. Large peaks appear at nouns and punctuation.

5.2 Semantic patterns

We measure the ratio of attention mass on named entities (NE) to quantify semantic specialization (Figure 3b). The KL-divergence between document and attention NE-distributions is not useful because, unlike POS tags, not every token has a NE tag. We find that for some attention heads, this ratio on average is more than thrice the ratio of named entity tokens in articles (0.3 and 0.1, respectively). The decoder head with the highest ratio lines up with the 'location head' we found in our visualizations. However, even though 'semantic' heads specialize, they still place large amounts of attention on other tokens. This appears to be due to the softmax that guarantees a smooth attention distribution. This is not necessarily desirable for interpretability purposes and makes reasoning about the roles of attention heads difficult.

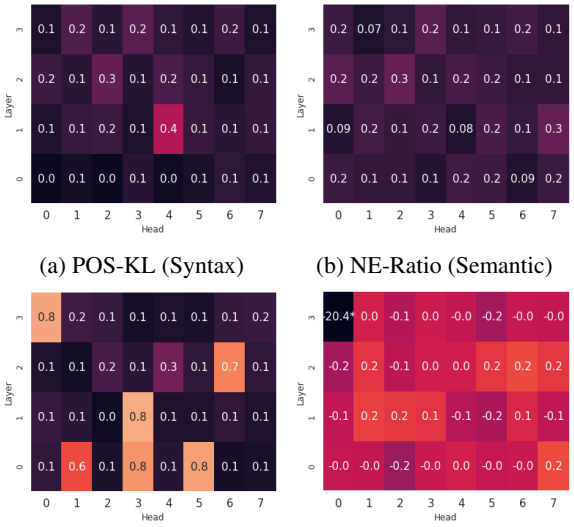
5.3 Positional patterns

We measure the ratio of each head's maximum attention weight per time step assigned to neighboring tokens (NE-ratio) and find six heads that consistently focus on preceding or succeeding tokens with a ratio of 0.8 or higher (Figure 3c). We find at least five decoder heads that focus on preceding, succeeding or currently generated tokens with a ratio of 0.7 or higher, even though there is no explicit supervisory 'copy mechanism' signal. This behavior brings to mind the inductive bias in (Bi-)RNN architectures where tokens are explicitly processed sequentially. The transformer, which does not have such inductive bias and solely uses attention, learns a similar way of processing. The positional activations are much higher compared to syntactic and semantic activations. This could imply that a focus on relative position is the most important specialization for abstractive summarization, or simply that its an easier task for attention heads to learn.

5.4 Does initialization affect specialization?

We train two identical models with different random seeds. We find similarities as well as slight differences in specialization. Both models learn a similar number of relative position heads, but the second model does not contain a head that focuses on locations. This is interesting because this was exactly a head that we deemed important to the summarization process in Section 5.

Perhaps one model outperforms the other in



(a) POS-KL (Syntax) (b) NE-Ratio (Semantic)

(c) Relative Position (d) Ablation

Figure 3: Metric activations for baseline decoder heads. An asterisk in (d) depicts statistical significance for the model with ablated head with respect to the non-ablated model (t-test on 1,000 summaries).

terms of correctly including locations in its predicted summaries, but is worse in terms of grammar. We measure the per-document ROUGE scores of both models on 1K articles and compute the differences between them. We find an average difference of eight ROUGE points. We hypothesize that this at least in part explains the difference in syntactic and semantic specialization between models: both models have different strengths and weaknesses. These findings show that we should be careful with interpreting attention heads – even if we could conclude that a particular instance of a trained model focuses on certain interpretable metrics, this is no guarantee that the model will always focus on this.

5.5 Ablating heads

To further investigate the importance of attention heads and their actual impact on the resulting summaries, we perform an ablation study. Following Michel et al. (2019), we add a binary gate to each attention head that allows us to exclude information flow from individual attention heads. Interestingly, we find that not a single ablated attention head causes a statistically significant different ROUGE-1 score. [3] The difference in ROUGE-1 after ablating an individual head is shown in Figure 3d. Significance is depicted with an asterisk. This demonstrates that attention heads, even those that seem to perform an interpretable task, do not affect model performance. It is a strong indicator that one should be careful in using the attention mechanism for transparency in abstractive summarization.

[3] Except for the copy-head, which is jointly trained to copy tokens directly from article to summary.

6 Improving Transparency of MHA

So far, we have found that it is unclear what MHA focuses on, for three reasons: (i) multiple heads focus on similar patterns, (ii) specialized heads still assign a considerable portion of their attention mass to tokens outside their specialization, and (iii) individual heads can be shut down without affecting the model performance.

We consider a method that was recently proposed to improve MHA in terms of interpretability, adaptive sparsity (Section 6.2), for our task of abstractive summarization. We then apply pruning to evaluate the resulting model in terms of redundancy, as well as our specialization metrics and ablation study introduced in Section 5. We start by applying pruning to our baseline model to investigate the effects on abstractive summarization.

6.1 Pruning attention heads

To further investigate redundancy in MHA we encourage the model to freely prune attention heads. We adopt a pruning strategy proposed by Voita et al. (2019). Our motivation for using this method is twofold: First, we want to know if we observe a similar number of redundant heads for the task of abstractive summarization compared to NMT, and how removing these heads affects specialization. This is interesting since attention was designed for NMT with an intuitive meaning: alignment. In abstractive summarization, however, the meaning of attention is less obvious. Second, we expect pruning to provide additional insights into the importance of attention heads.

To remain consistent with our previous analysis, we prune encoder self-attention and decoder cross-attention heads. Following Voita et al. (2019), we add a gate to each attention head and consider it a trainable parameter, unlike the binary gates used for ablation. We then encourage the model to set these gates to exactly 0 with a stochastic relaxation of the L_0 regularization penalty from Eq (2)) to the summarization loss:

$$\text{loss} = \frac{1}{T}\sum_{t=1}^{T} -\log P_v(w_t^*) + \lambda L_C(\phi). \quad (4)$$

The first term is the cross-entropy, P_v is the pre-

λ	#Pruned (enc/dec)	R-1 F1	R-2 F1	R-L F1
0		38.76	17.13	36.00
1	2/0	39.12	17.15	36.24
3	20/14	38.67	16.66	36.06

Table 2: ROUGE scores after pruning. *#0-G* shows the number of exactly-zero gates for encoder self-attention and decoder cross-attention, respectively.

dicted distribution over the extended vocabulary, T is the number of decoding time steps (predicted tokens), and w_t^* is the target token for time step t. We test several λ's to find an optimal value that prunes the largest amount of heads without decreasing performance.

6.1.1 Results

We are able to prune 34 out of 64 attention heads without a large performance impact (Table 2). For some values of λ, the pruned model actually outperforms the baseline models (Table 1). We believe this to be caused by the regularizing nature of the L_C norm, which reduces the number of parameters and causes stronger generalization. This confirms the hypothesis that many heads are redundant, and is in line with results from our ablation study as well as results from Voita et al. (2019). It is another strong indicator that attention heads should not be used for transparency.

We observe that all relative position heads have been retained, comprising almost half of the remaining heads. This is interesting; these heads are interpretable, but they do not have semantic meaning for the task of summarization. Syntactic pattern activations, measured with POS-KL, are lower compared to the baseline. Semantic pattern activations, measured with NE-Ratio, are of similar magnitude. Interestingly, the interpretable 'location' head was also pruned.

When ablating individual heads on the pruned model, we observe six heads that cause a statistically significant drop in ROUGE. This indicates that the remaining heads are more important to model performance compared to heads in the baseline models (Section 3d). We also observe that these heads mostly correspond to strong relative position heads. It thus appears that relative position is in fact the most important specialization to summarization performance. These findings are another piece of evidence for redundancy. More importantly, even though heads appear to be interpretable (such as the 'location' head), they do not necessarily contribute to the overall summarization performance. Thus, one should be careful with using them for transparency. In the next section we investigate a method that was recently introduced to increase the interpretability of multi-head attention in NMT.

6.2 Sparse attention

Correia et al. (2019) propose to replace the softmax activation function with α-entmax to improve the interpretability of multi-head attention. In an attempt to improve the interpretability even further, we use the sparsest case of α-entmax instead: the sparsemax transformation.

In Section 5 we observed that specialized heads place a considerable amount of attention mass on non-related tokens. The sparsemax activation function seems well suited to address this problem. Sparsemax projects an input vector $\mathbf{z}$ onto the probability simplex and is formally defined as:

$$\text{sparsemax}(\mathbf{z}) = \underset{\mathbf{p}\in\Delta^{K-1}}{\text{argmin}} ||\mathbf{p}-\mathbf{z}||^2 \quad (5)$$

$$\Delta^{K-1} = \{\mathbf{p}\in\mathbb{R}^K \mid \mathbf{1}^T\,\mathbf{p} = 1, \geq \mathbf{0}\}, \quad (6)$$

resulting in the following attention function:

$$\text{Attn}(Q,K,V) = \text{sparsemax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (7)$$

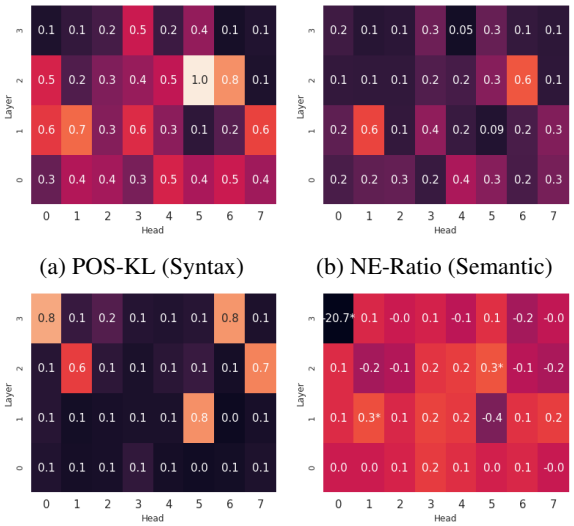
We first apply this modification to the **Enc**oder (*Sparse-Enc* model). We then extend it to the entire model. However, we observe that we need to exclude the **T**op **L**ayer of the decoder (*Sparse-TL* model) to prevent performance from collapsing. We hypothesize that this is due to interference with the copy-head, located in the top layer. To further test this hypothesis, we individually exclude the **C**opy **H**ead (*Sparse-CH* model) instead of the entire top layer.

6.2.1 Results

We can push 97% of all attention weights to zero for all sparsemax heads in the encoder without affecting performance. Sparse-TL and Sparse-CH both perform competitively with a minor drop in ROUGE (Table 3). This is in line with findings in NMT by Correia et al. (2019), although their α-entmax model is free to choose the degree of sparsity, unlike ours. We find that the copy head is indeed the cause of performance collapse as applying sparsemax results in roughly half the ROUGE score. We hypothesize that the sparsemax activation interferes with the loss computation or OOV

	R-1 F1	R-2 F1	R-L F1
Sparse-Enc	38.73	16.68	35.64
Sparse-TL	38.51	16.87	35.66
Sparse-CH	38.30	16.47	35.35

Table 3: ROUGE scores for using sparsemax on all **Enc**oder heads, all heads except for the decoder **T**op **L**ayer, and the entire model except for the **C**opy **H**ead.



(a) POS-KL (Syntax) (b) NE-Ratio (Semantic)

(c) Relative Position (d) Ablation

Figure 4: Metric activation for Sparse-TL decoder heads.

token sampling of the copy head, but want to investigate this in more detail.

Figure 4 shows stronger activation on syntactic (POS-KL) and semantic (NE-Ratio) patterns compared to the baseline (Figure 3). Upon closer inspection of heads with a high syntactic activation we see significantly more peaked distributions over POS tags (Figure 5 shows three encoder heads with the highest POS-KL). Interestingly, an unseen strong syntactic pattern emerges that focuses on determinants or pronouns. This again stresses that different specializations can emerge in different models.

Do stronger activations on our metrics imply that individual heads have become more important to summarization? We observe nine heads that statistically significantly affect performance when ablated. It thus seems that not only sparsemax heads specialize more distinctly, but that the impact of individual heads has also increased.

6.3 Pruning sparse attention heads

We have discovered that we can prune roughly half the attention heads and that a sparsemax activation function appears to improve transparency. If we now prune a transformer model with sparse attention, we would expect fewer heads that can be pruned. After all, the sparsemax-activated heads appear more interpretable as well as faithful.

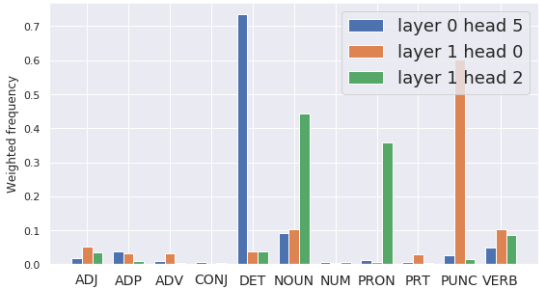


Figure 5: Highly focused attention distributions over POS tags. We show the top three specialized syntactic encoder heads in the Sparse-TL model.

We prune the Sparse-TL model of which the top decoder layer heads retain their softmax because (i) its performance is superior, and (ii) we want to investigate whether there is a preference for pruning heads with either a softmax or a sparsemax activation function.

6.3.1 Results

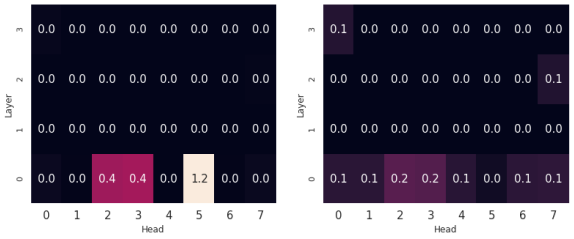
Surprisingly, we are able to prune an even larger amount of heads in our sparsemax model compared to the baseline model: 43 out of 64 heads (Table 4). All heads in the decoder top layer are retained. Perhaps the freedom that a softmax activation provides by allowing for more diffuse distributions is important, or the heads in the top decoder layer were incidentally more important.

#Pruned (enc/dec)	R-1 F1	R-2 F1	R-L F1
22/21	38.45	16.63	35.39

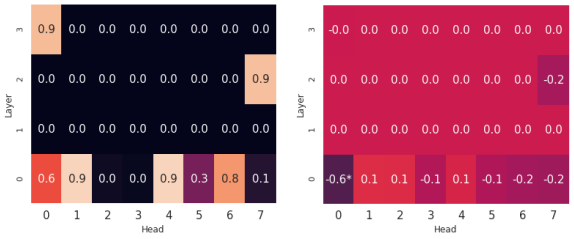
Table 4: Results for a sparse transformer (Sparse-TL) after pruning on ROUGE. λ is emperically set to 2.

We find that five out of the ten remaining encoder heads focus strongly on relative positions (Figure 6). The remaining encoder heads score high on syntactic patterns. However, upon inspection of the strongest syntactic encoder head we discover it that almost exclusively focuses on the word 'the'. This does not appear to be a relevant specialization to summarization, but nonetheless this head is one of the few encoder heads that survived the pruning process.

Similar to the encoder, the majority of decoder heads focus on relative position (Figure 7). The others respond slightly higher to semantic as well as syntactic metrics compared to the baseline, but not by a large margin. For heads in the top layer

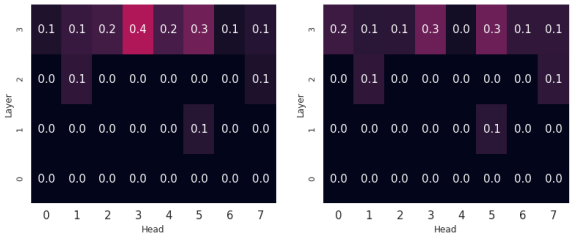


(a) POS-KL (Syntax) (b) NE-Ratio (Semantic)

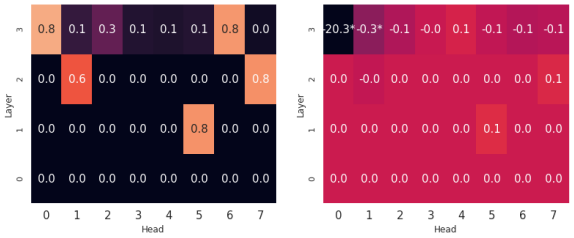


(c) Relative Position (d) Ablation

Figure 6: Metric activation for Sparse-TL encoder heads after pruning.



(a) POS-KL (Syntax) (b) NE-Ratio (Semantic)



(c) Relative Position (d) Ablation

Figure 7: Metric activation for Sparse-TL decoder heads after pruning.

this could be expected, as they use the same softmax activation function as heads in the baseline.

We can conclude that sparsemax appears to improve the interpretability as well as faithfulness of MHA. It scores high on our specialization metrics, and contains more heads with a statistically significant impact on performance. However, when we prune a sparsemax model, we are able to prune even more heads than we could in the baseline model. Additionally, most of the semantic and syntactic specialized heads disappear. The remaining heads are predominantly relative position heads. This gives rise to an important question: does sparsity in multi-head attention actually improve transparency?

7 Discussion

How do we evaluate whether attention can be used as means for transparency? This question is raised over and over again, and is very difficult to answer. This is illustrated by a large body of recent work that adresses this question (see Section 3). Quantifying specialization in attention heads, pruning and ablation studies provide more insights, but still result in contradictory observations. A promising recent line of work focuses on adversarial attention attacks (Jain and Wallace, 2019; Serrano and Smith, 2019; Vashishth et al., 2019), but this is not yet applied to sequence-to-sequence tasks. We believe this to be a promising next step in better understanding attention as means for transparency. Finally, in this work, as well as in most related work, we assume that representations learned within a transformer model correspond to a (contextualized representation of) the input token at that position. This assumption should be properly investigated.

8 Conclusion

We have investigated to what extent multi-head attention in abstractive summarization is transparent. We have introduced quantitative metrics that showed that multi-head attention is partially interpretable. However, we have also shown that all individual heads can be ablated without a significant drop in performance, indicating that one should be very careful using the attention mechanism for transparency in abstractive summarization. Replacing the softmax activation function by a sparsemax activation function resulted in improved scores on our interpretability metrics, and fewer heads that could be ablated without decreasing summarization performance. However, in this setting more heads can be pruned. In all our experiments we find that relative position heads seem integral to performance and persistently remain after pruning.

Taking all our findings and related work into account, we believe that for multi-head attention to be transparent, it should adhere to the following criteria: (i) multi-head attention should have a minimum number of heads, (ii) heads should have no overlap in specialization but focus on distinct representational subspaces, and (iii) we need the right metrics to measure interpretability.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Joris Baan, Maartje ter Hoeve, Marlies van der Wees, Anne Schuth, and Maarten de Rijke. 2019. Do transformer attention heads provide transparency in abstractive summarization? *arXiv preprint arXiv:1907.00570*.

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

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.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Gonçalo M Correia, Vlad Niculae, and André FT Martins. 2019. Adaptively sparse transformers. *arXiv preprint arXiv:1909.00015*.

Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. 2018. Latent alignment and variational attention. In *Advances in Neural Information Processing Systems*, pages 9712–9724.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. *arXiv preprint arXiv:1808.10792*.

Reza Ghaeini, Xiaoli Z Fern, and Prasad Tadepalli. 2018. Interpreting recurrent and attention-based neural models: a case study on natural language inference. *arXiv preprint arXiv:1808.03894*.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An approach to evaluating interpretability of machine learning. *arXiv preprint arXiv:1806.00069*.

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

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Christos Louizos, Max Welling, and Diederik P Kingma. 2017. Learning sparse neural networks through l_0 regularization. *arXiv preprint arXiv:1712.01312*.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Chaitanya Malaviya, Pedro Ferreira, and André F. T. Martins. 2018. Sparse and constrained attention for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 370–376, Melbourne, Australia. Association for Computational Linguistics.

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

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *arXiv preprint arXiv:1905.10650*.

Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Vlad Niculae, André FT Martins, Mathieu Blondel, and Claire Cardie. 2018. Sparsemap: Differentiable sparse structured inference. *arXiv preprint arXiv:1802.04223*.

Alessandro Raganato, Jörg Tiedemann, et al. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*.

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.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*.

Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*.