UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Do Transformer Attention Heads Provide Transparency in Abstractive Summarization?

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

Publication date
2019

Document Version
Author accepted manuscript

Published in
Proceedings of FACTS-IR 2019

Link to publication

Citation for published version (APA):
Baan, J., ter Hoeve, M., van der Wees, M., Schuth, A., & de Rijke, M. (2019). Do Transformer Attention Heads Provide Transparency in Abstractive Summarization? In *Proceedings of FACTS-IR 2019* ArXiv. https://arxiv.org/abs/1907.00570

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:10 Nov 2026

Do Transformer Attention Heads Provide Transparency in Abstractive Summarization?

Joris Baan[1, 2] Maartje ter Hoeve[1] Marlies van der Wees[2] Anne Schuth[2] Maarten de Rijke[1]
[1]University of Amsterdam, Amsterdam [2]De Persgroep, Amsterdam
joris.baan@student.uva.nl, m.a.terhoeve@uva.nl, marlies.van.der.wees@persgroep.nl,
anne.schuth@persgroep.nl, derijke@uva.nl

arXiv:1907.00570v2 [cs.CL] 8 Jul 2019

ABSTRACT

Learning algorithms become more powerful, often at the cost of increased complexity. In response, the demand for algorithms to be transparent is growing. In NLP tasks, attention distributions learned by attention-based deep learning models are used to gain insights in the models' behavior. To which extent is this perspective valid for all NLP tasks? We investigate whether distributions calculated by different attention heads in a transformer architecture can be used to improve transparency in the task of abstractive summarization. To this end, we present both a qualitative and quantitative analysis to investigate the behavior of the attention heads. We show that some attention heads indeed specialize towards syntactically and semantically distinct input. We propose an approach to evaluate to which extent the Transformer model relies on specifically learned attention distributions. We also discuss what this implies for using attention distributions as a means of transparency.

CCS CONCEPTS

• **Information systems** → *Summarization.*

KEYWORDS

Transformer, Abstractive summarization, Attention, Transparency

1 INTRODUCTION

When trusting a machine-generated summary it may be crucial to have an understanding of how this summary came to be. Attention mechanisms [2, 12] have gained popularity in the context of deep learning-based approaches to summarization [15, 19, 20]. Briefly, classic attention mechanisms learn a function that assigns a score to each encoder's hidden state based on its relevancy to the word being decoded. Through a weighted average with these softmaxed scores, hidden states with high scores are amplified. Because they provide an interpretable heatmap over an input sequence, attention mechanisms have been used to gain insights in the behavior of a given model [3, 9, 11]. However, this may be misleading. First, in commonly used architectures such as Bidirectional Recurrent Neural Networks (Bi-RNNs) [18] and Transformers [20] a lot of computation takes place between an input token and the hidden state. As a result, it is unclear whether the hidden state that an attention weight operates on corresponds to its input token. Second, shown heatmaps are usually cherry picked and do not necessarily generalize over all examples [3, 9, 11]. Third, different attention distributions can lead to the same model output, which implies that "attention is not explanation" [9].

Can attention distributions in a Transformer model [20] trained for abstractive summarization be used to address model transparency for the summarization task? Transformer models consist of a modular, multi-headed structure. Because of this modularity, we may be able to find distinct interpretable patterns that generalize over a large number of examples.

We adopt a qualitative and a quantitative approach to investigate the behavior of attention heads. For the qualitative approach we visually inspect the encoder self-attention and decoder cross-attention and find that some heads attend to locations, persons, organization nouns, or punctuation. We then introduce several metrics to quantitatively evaluate whether the previous findings generalize over 1K news articles as well as different initializations of the model. Importantly, by doing so we move away from cherry picked attention heatmaps. We then discuss a method to investigate to what extent the Transformer model relies on certain attention distributions, inspired by recent work on adversarial attention [9]. This raises the question whether adversarial methods invalidate the use of learned attention distribution as a means for transparency.

With this work we contribute: (1) quantitative metrics that measure the degree to which attention heads specialize towards attending Part-of-Speech (POS), Named Entity (NE) tags and relative position; and (2) a method for adversarial attacks on seq2seq Transformers to assess the effect of individual attention heads on model output.

2 RELATED WORK

Transparency in machine learning has become important as models become more complex and more frequently play a role in decision making [7, 14]. Terms such as explainability and transparency are hard to define and open for multiple interpretations. Gilpin et al. [7] describe an *explanation* to be an answer to "why questions" and consider it a trade-off between *interpretability* and *completeness*. *Interpretability* means being understandable to humans, whereas *completeness* covers how well the explanation is faithful to the actual model mechanics. Doshi-Velez and Kim [5] note that interpretability can be used to evaluate desiderata besides performance such as causality or trust. Mittelstadt et al. [14] argue that *transparency* addresses how a model functions internally. Such a model or its components can be called *transparent* when they can be comprehended entirely. Following [5, 14], we maintain that a fully transparent

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Paris '19, June 21–25, 2019, Paris, France
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

model should be understandable for a user. However, since fully transparent models are not always capable of competitive performance, we argue that a step in the direction of a more interpretable model already provides transparency. We want to understand the attention mechanism and its role in the Transformer to assist in the discussion on whether attention provides transparency.

2.1 Attention for transparency in NLP

The following recent work is aimed towards a better understanding of attention distributions and whether it can be used to explain a model. Raganato et al. [17] study the self-attention of a Transformer encoder for NMT and observe that some heads mark syntactic dependency relations. Vig [21] visualizes BERT's [4] self-attention and finds patterns such as attention to the surrounding words, identical/related words, predictive words and delimiter tokens. Concurrent to our work, Voita et al. [22] perform a similar analysis of multi-headed attention in NMT and Michel et al. [13] for BERT [4]. They find that heads specialize towards linguistically interpretable roles, but that a majority can be pruned after training without affecting performance. Jain and Wallace [9] observe that attention is commonly (implicitly or explicitly) claimed to provide insight into model dynamics. They argue that if attention is used as explanation, it should exhibit two properties: (1) attention should correlate with feature importance measures; and (2) adversarially crafted attention distributions should lead to different predictions, or be considered equally plausible explanations. Such an adversarial attention distribution should maximally differ from the learned attention, while the corresponding output distribution is constrained to be the same within a small range ϵ. With a Bi-RNN or CNN encoder it is possible to construct such adversarial distributions for NLP classification tasks such as binary text classification. They argue that attention heatmaps should thus not be so easily assumed to provide transparency for model predictions [9].

2.2 The Transformer

The Transformer is a seq2seq model that relies solely on (self-)attention and stacks several encoders and decoders. *Self-attention* computes scores between each of the input tokens, as opposed to computing scores between encoder and decoder hidden states, referred to as *cross-attention*. *Multi-headed* attention refers to having multiple "representation subspaces" or *heads* governed by separate sets of W_Q, W_K, W_V weight matrices. These matrices project each input into a query, key and value vector from which scores and context vectors are computed. The attention function itself is *scaled dot product attention* and identical to Luong et al. [12]'s dot-product attention apart from the scaling factor (Eq. (2)). H represents an embedding for the bottom encoder/decoder and a hidden state for the remainder (Eq. (1)).

$$head_i = Attention(HW_Q^i, HW_K^i, HW_V^i) \tag{1}$$

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}V}\right). \tag{2}$$

Our work extends and differs from the related work just discussed in three important ways: (1) we introduce three metrics relevant to summarization to quantify patterns in attention; (2) we analyze the decoder cross-attention in addition to the encoder self-attention; and (3) our input sequences (news articles) are significantly longer than the short sentences used in previous work.

structure an agreement or series of agreements so that Iran could not covertly develop a nuclear arsenal before the United States and its allies could respond . The new framework has exceeded expectations in achieving that goal . It would reduce Iran 's low-enriched uranium stockpile , cut by two-thirds its number of installed centrifuges and implement a rigorous inspection regime . Another dubious assumption of opponents is that the Iranian nuclear program is a covert weapons program . Despite sharp accusations by some in the United States and its allies , Iran denies having such a program , and U.S. intelligence contends that Iran has not yet made the decision to build a nuclear weapon . Iran 's continued cooperation with International Atomic Energy Agency inspections is further evidence on this point , and we 'll know even more about Iran 's program in the coming months and years because of the deal . In fact , the inspections provisions that are part of

Figure 1: Attention head focusing on locations.

3 EXPERIMENTAL SETUP

We adopt OpenNMT's implementation [10] of the CopyGenerator Transformer [6]. Both encoder and decoder have four layers with eight heads. We use scaled dot attention, Gehrmann et al. [6]'s new summary specific coverage function, Wu et al. [23]'s length penalty during beam-search decoding at inference time, and See et al. [19]'s pointer generator architecture.

We use the *CNN/Daily Mail* [8, 15] dataset containing roughly 300.000 news articles and use the script provided by Nallapati et al. [15] to split this into a train, test and validation set. Articles consist on average of 781 tokens and summaries of 56 tokens. Following See et al. [19] we truncate articles to 400 words. We train two identical models with different parameter initializations to investigate whether stochasticity affects the way attention heads specialize. Both models have similar ROUGE scores (**ROUGE-1**: 38.76/38.81, **ROUGE-2**: 17.13/16.77, **ROUGE-L**: 36.00/36.28).

4 QUALITATIVE APPROACH

We extend a tool originally created to visualize a copy-generator model by See et al. [19]. It highlights words in an input article based on the magnitude of their corresponding attention weights and gives control over which attention type, layer or head to visualize.[1] We compute an overall attention distribution by summing and normalizing attention weights over all time steps.

For the encoder, the vast majority of the attention heads seem to focus on preceding, succeeding or surrounding words. Similarly for the decoder, several heads find an occurrence of the currently or previously decoded word. Some heads seem to focus on punctuation and delimiters overall, confirming observations from Vig [21].

Strikingly, when inspecting the overall decoder attention, there are heads that seem to focus on key words, locations (Figure 1), organizations, people or days of the week. These heads appear to have learned to detect such entities without explicit supervisory signals. However, there are plenty of articles for which these patterns are less obvious (Figure 2). Such "counter examples" might indicate that these patterns do not generalize and are based on our bias for interpretability, or the model might sometimes fail to predict the specialized attention, similar to how the ROUGE score is lower for some documents than others.

5 QUANTITATIVE APPROACH

To support our findings from the qualitative visualizations and examine to what extent the observation generalize, we introduce three quantitative metrics.

[1] Our version is publicly available at https://aijoris.github.io/attnvis/.

Hong Kong (CNN) There 's a booming black market in Hong Kong , but it 's not for fake Apple Watches , or the iPhone . Instead , people are going crazy for tins of butter cookies . Tourists and locals line up around the block for several hours just to get their hands on Jenny 's cookies -- at $ 9 a tin . Its popularity has spurred bakeries to make and sell knockoffs , and the original store has signs warning against buying 'fake ' Jenny 's cookies . The tiny shop , located in Tsim Sha Tsui , one of the city 's main shopping districts , is swarming with people handing over wads of cash for the `` little bear cookies '' as they are known across Asia . People are even hired to stand in line to buy the goods and are later resold at a 70 % mark-up yards away , something the bakery also tries to discourage . A few meters away from the long cookie line , old ladies hold up paper signs advertising the cookies for sale . But when they see cameras approaching , they scurry away , only to reappear on another street corner . The frenzy in Hong Kong over the buttery treats is by no means an isolated example . In other parts of the world , food mania has erupted , swiftly winning people 's hearts and stomachs , only to fizzle out in a few months . From cronuts to ramen burgers , here are some foods that people around the world have spent hours of their lives waiting for . Were they worth it ?

Figure 2: Attention head that seemed to focus on named entities fails to do so in the above example.

5.1 Relative position

We record how often the maximum attention weight is on preceding or successive tokens relative to the token currently being encoded or decoded. Figure 3 shows that at least nine encoder heads and four decoder heads focus on relative positions. This behavior brings

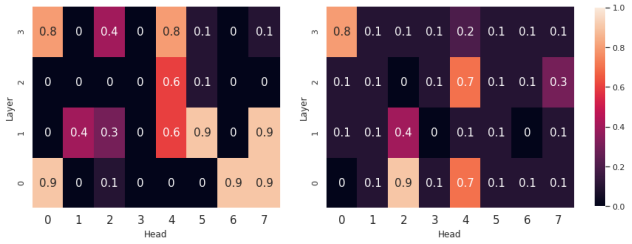


Figure 3: Ratio of the max attention weight being assigned to neighboring tokens. (Left): Encoder. (Right): Decoder.

to mind the inductive bias in an RNN where tokens are explicitly processed sequentially, or a CNN using convolutions to construct hidden states. The Transformer does not have such inductive bias and solely uses attention. Interestingly, some heads appear to have learned a similar way of processing nonetheless. Model 2 is not shown, but has a similar number of relative position heads.

5.2 POS-KL

We tag each article in the test set (see Section 3) with 12 universal part-of-speech tags [16] using a POS tagger by FLAIR [1]. For every article we compute a histogram of POS tag counts to serve as the baseline. Then, for each head these tag counts are multiplied by the accumulated attention weights of all tokens labeled with that tag and normalized. The degree to which an attention head is specialized can be measured by the difference between its attention-weighted POS tag distribution and the baseline POS tag distribution. We use the KL-divergence to quantify this difference and average these over all articles.

Figure 4a shows the decoder cross-attention weighted POS tag distributions for three heads with the highest KL-divergence. The peaks at the punctuation, noun and verb tag confirm that some heads consistently focus on specific word categories. For the two trained models, different specializations emerge. Model 2 has two heads with a large peak for verbs, and all three heads have a relatively high peak at punctuation as well. Model 1 has no such peaks for verbs and only one of the top three heads that focuses on punctuation.

5.3 NEP

Each article is tagged with four named entities: *persons*, *locations*, *organizations*, and *miscellaneous*, using a NE tagger by FLAIR [1]. Unlike POS tags, however, not each token is a named entity. This can cause a high KL-divergence between the attention weighted named entity distribution and baseline (NE-KL), even if a head barely attends to named entities. We found computing the proportion of attention mass over all named entities (NEP) to be a better method for detecting specialized heads.

The baseline ratio of named entities over articles is 0.1. Figure 4b shows the top three cross-attention weighted distributions over named entities based on NEP. Heads shown have a NEP of at least double the baseline ratio. Large peaks at persons and organizations can be observed for both models. Model 1's most specialized head corresponds to the 'location head' found in our qualitative analysis. This indicates the ability to detect specialized heads using NEP. It simultaneously provides more insight into what such a head actually attends and how well our qualitative findings generalize. We refer the reader to the appendix for a complete overview of the metrics for all attention heads, including standard deviations.

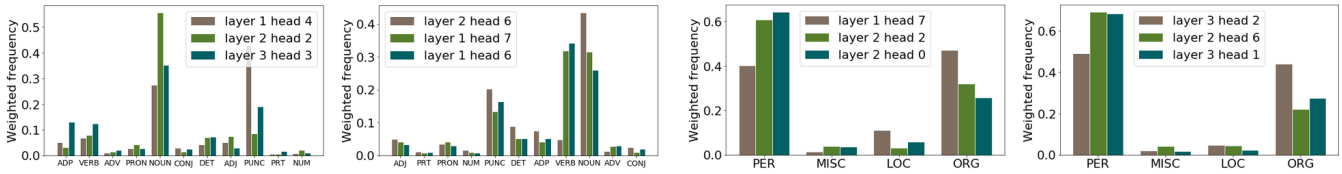
5.4 Analysis

We did not detect any POS or NE specialization for the encoder's overall self-attention. This is in line with the earlier observation that most encoder heads attend relative positions. It is important to note that we have not evaluated the models on per-document ROUGE scores. This could explain the observed difference in specialization between models. Perhaps model 2 performs better on articles for which verbs are important in the summary, resulting in a head that more explicitly attends verbs. Another note is that not every article contains named entities, causing a decrease in NEP. One interesting example can be found in Appendix A, where a NE-specialized head highlights lions in one article. Lions are not named entities but do fulfil a similar role, indicating that NEP might not always fully capture a specialization.

The main takeaway is that we show that some attention heads specialize towards attending relative locations, nouns, verbs, punctuation, persons, locations or named entities. The top 3 specialized heads that were found using our quantitative approach line up with findings from visualizations. However, an analysis of POS-KL and NEP distributions over articles also indicate that heads only specialize to some extend and sometimes take into account a considerable amount of non-related tokens. This supports claims by Jain and Wallace [9], urging the research community to be careful in using attention as explanation.

6 ADVERSARIAL ATTENTION

Given that some attention heads are found to focus on interpretable input, we want to understand to what extent the model actually relies on these specific attention distributions. For future work, we propose to adapt the adversarial attention method by [9] to make it compatible with a seq2seq Transformer model using beam search. Instead of requiring the output distributions to be within a small ϵ, it is sufficient to constrain the top K output probabilities of each decoding to be within a small ϵ, whereby $K = beamsize$. This will result in identical output sequences, since the beam search path with the highest probability remains the same. As a consequence, we can craft one adversarial attention distribution for each decoding step and aggregate them to evaluate the overall success on a summary.



(a) Based on POS-KL. (Left): Model 1; (Right): Model 2. **(b) Based on NEP. (Left): Model 1; (Right): Model 2.**

Figure 4: A comparison of the top 3 specialized heads.

Additionally, we propose to modify the attention distribution of a specialized head to attend another specific phenomenon. For example, we could construct a distribution that solely attends persons for a head that specializes towards locations and observe whether locations change into persons. The Transformer model is large, in our case containing 32 heads for both the encoder and decoder. It is unclear to what degree modifying the attention distribution of one head can be expected to affect the output summary.

However, if such an adversarial distribution can be constructed, it raises the question to what extend it invalidates the learned attention distribution as means for transparency. Should an attention distribution have a causal relationship with the model output in order to use it for transparency, or does the fact that the model has learned this distribution justify using it as such? Similarly, does the use of attention heads to address transparency become invalidated if different specializations form for architecturally identical model on the same data set? Or does this add to its value because it shows differences between models that otherwise remain undetected?

7 CONCLUSION

We have presented a qualitative and quantitative approach to better understand what Transformer attention heads attend to in abstractive summarization. Some attention heads do specialize towards interpretable parts of a document, but this does not apply to all documents. We confirm this with three proposed metrics that quantify what heads focus on in terms of POS tags, named entities and relative position. We also find that these specializations are not consistent over differently initialized models. Finally, we discuss the use of adversarial attention to examine the effect of attention distributions on model output, and ask what such adversarial methods imply for transparency.

One limitation of this work is that there is no proof that the index of a hidden state corresponds to a (contextual) representation of the corresponding input token. A natural question is why specialized heads perform poorly on some articles. Future work could compare per-document ROUGE with POS-KL and NEP to examine correlations between summarization and head specialization.

ACKNOWLEDGMENTS

We thank Mostafa Dehghani for valuable discussions and feedback. This research was partially supported by Ahold Delhaize, the Association of Universities in the Netherlands (VSNU), the Innovation Center for Artificial Intelligence (ICAI), and the Police AI lab. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*. 1638–1649.
[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473* (2014).
[3] 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*. 3504–3512.
[4] 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* (2018).
[5] Finale Doshi-Velez and Been Kim. 2017. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608* (2017).
[6] Sebastian Gehrmann, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up Abstractive Summarization. *arXiv preprint arXiv:1808.10792* (2018).
[7] 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* (2018).
[8] 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*. 1693–1701.
[9] Sarthak Jain and Byron C Wallace. 2019. Attention is not Explanation. *arXiv preprint arXiv:1902.10186* (2019).
[10] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1701.02810* (2017).
[11] Tao Lei. 2017. *Interpretable Neural Models for Natural Language Processing*. Ph.D. Dissertation. Massachusetts Institute of Technology.
[12] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. *arXiv preprint arXiv:1508.04025* (2015).
[13] Paul Michel, Omer Levy, and Graham Neubig. 2019. Are Sixteen Heads Really Better than One? *arXiv preprint arXiv:1905.10650* (2019).
[14] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2018. Explaining Explanations in AI. *arXiv preprint arXiv:1811.01439* (2018).
[15] 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* (2016).
[16] Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A Universal Part-of-Speech Tagset. *arXiv preprint arXiv:1104.2086* (2011).
[17] Alessandro Raganato, Jörg Tiedemann, et al. 2018. An Analysis of Encoder Representations in Transformer-Based Machine Translation. In *2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. ACL.
[18] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing* 45, 11 (1997), 2673–2681.
[19] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the Point: Summarization with Pointer-Generator Networks. *arXiv preprint arXiv:1704.04368* (2017).
[20] 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*. 5998–6008.
[21] Jesse Vig. 2018. Deconstructing BERT: Distilling 6 Patterns from 100 Million Parameters. *towardsdatascience.com/deconstructing-bert-distilling-6-patterns-from-100-million-parameters-b49113672f77*. Accessed: 2019-04-29.
[22] 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* (2019).
[23] Yonghui Wu et al. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144* (2016).

A APPENDIX

Roma have been ordered to close part of their stadium for their next home game after supporters were sanctioned for showing an offensive banner over the weekend . The banner , displayed during Roma 's 1-0 defeat of Napoli on Saturday , caused outcry for insulting the mother of a Napoli fan who was killed in violent clashes between supporters last year . A Serie A disciplinary tribunal stated the banners were 'by their content provocatively insulting to the mother of a supporter of the opposing team , who died in dramatic circumstances ' . Roma players celebrate their 1-0 victory over Napoli , but the club have been punished for an offensive banner Roma fans display banners during the game , although a banner (not shown) taunting the mother of a Napoli fan who died after violent clashes last year caused outcry Roma will have part of their stadium closed for the league match with Atalanta on April 19 The taunts were aimed at Antonella Leardi whose son Ciro Esposito died after being shot during violent clashes that followed last year 's Coppa Italia final in Rome . Although Napoli defeated Fiorentina 3-1 in the final , but the incident involved clashes with Roma supporters - encouraging Leardi to start a campaign against football violence . Roma 's Stadio Olimpico Stadium will be part closed for game against Atalanta on April 19 . A young Napoli fan died after clashes between supporters following last year 's Coppa Italia final Napoli defeated Fiorentina in the 2014 Coppa final , but Roma fans were involved in clashes after the game Club president James Pallotta has condemned the club 's fans in a statement on their official website : 'As has been expressed repeatedly , AS Roma considers any events that lead to the loss of life at a football match to be a defeat for civil society as a whole , regardless of affiliations to clubs or fan groups . 'The enormous pain that follows such events deserves maximum , unconditional respect from all and necessitates that all parties -- fans , clubs and law enforcement agencies -- strive to ensure that such pain is not renewed , not even verbally , in the stands of a stadium . ' The mother , Leardi , told the ANSA news agency : 'May God change the hearts of those people

Figure 5: Specialized named entity head focusing on football teams.

Until now , it has been a hidden world , protected by hundreds of feet of freezing water and 20 metres of ice . Now , researchers have created a special robo-explorer to borrow into the ice and record the first footage of what lies on the seabed below the Ross Ice Shelf . The icefin was deployed (and retrieved) the vehicle through a 12-inch diameter hole through 20 meters of ice and another 500 meters of water to the sea floor . Scroll down for video One of the unique creatures spotted on the seabed.The icefin was deployed (and retrieved) the vehicle through a 12-inch diameter hole through 20 meters of ice and another 500 meters of water to the sea floor . A first-of-its-kind robotic vehicle recently dove to depths never before visited under Antarctica 's Ross Ice Shelf and brought back video of life on the seafloor . 'We built a vehicle that 's a hybrid between the really small probes and the ocean-going vessels , and we can deploy it through bore holes on Antarctica , ' said Britney Schmidt , an assistant professor in the School of Earth and Atmospheric Sciences at the Georgia Tech , and the principle investigator for the Icefin project . The technologies developed for Icefin will also help in the search for life on other planets , namely Europa , a moon of Jupiter . Antarctica 's icy oceans are remarkably similar to Europa'sice-capped oceans . 'At the same time , we 're advancing hypotheses that we need for Europa and understanding ocean systems here better . 'We 're also developing and getting comfortable with technologies that make polar science -- and eventually Europa science - more realistic . ' The robotic vehicle carried a scientific payload capable of measuring ocean conditions under the ice . Icefin 's readings of the environment under Antarctica 's ice shelves , and video of the life that thrives in these harsh conditions , will help understand how Antarctica 's ice shelves are changing under warming conditions , and to understand how organisms thrive in cold and light-free environments . The icefin was deployed (and retrieved) the vehicle through a 12-inch diameter hole through 20 meters of ice and another 500 meters of water to the sea floor . A team of scientists and engineers from the Georgia

Figure 6: Specialized head focusing on the location Antarctica.

A lion had an unfortunate accident when it got its head stuck in a feeding barrel at a zoo in the Netherlands . Captured by a visitor to the Dierenrijk zoo , the lion can be seen inserting its head into the barrel and attempting to retrieve a piece of meat . Zoo keepers place food in barrels to stimulate the lions , as it replicates the challenges faced when the animals feed in the wild . The animal , hoping to beat two other lions to the food , reaches in too far and suddenly gets its head stuck . Reacting in a panic the lion jumps backwards and attempts to flick the barrel from its head as the other lions chase after it . Visitors to the zoo can be heard laughing hysterically while the lion desperately tries to set itself free . The animal , hoping to beat two other lions to the food , reaches into the barrel too far and gets its head stuck The other lions chase after the barrel as the distressed animal runs around in a circle and attempts to free itself One lion appears to come to the rescue by jumping up and standing on the barrel while the lion pulls from the ground . But unfortunately the animal is unable to free its head and the other two lions walk around looking perplexed . Feeling sorry for itself , the distressed lion stands alone in the corner before turning and running frantically towards the zoo ' s visitors . The lion reacts in panic and attempts to flick the barrel from its head by running around Visitors watch on as the lion jumps around and even runs towards the people standing at the edge of the enclosure The video concludes with the lion looking well and truly fed up as it lies on the ground , exhausted by its failed attempts to break free . According to the video maker , the lion was eventually freed from the barrel and it was unharmed . He said : ' To make feeding a little bit more exciting for the lions they [the keepers] put meat in a barrel with two small openings . According to the video maker , the lion was eventually freed from the barrel and it was unharmed The other lions look perplexed and one even

Figure 7: Specialized NE head with a low NEP. This is interesting because this head attends animals in this article, which are not named entities. However, intuitively this example still shows a form of specialization, but this is not reflected by the NEP metric.

(CNN) As we approach April 27 when South Africa marks the anniversary of the first post-apartheid elections held that day in 1994 , we are faced with yet another wave of deadly attacks against African migrants . Outrage triggered by this violence is being heard loudly throughout social media with `` # WeAreAfrica '' showcasing the need for a common front against this affront . These recurring attacks against migrants and their property might be read as one more indication of how the rainbow nation 's dream has faltered . That vision not only symbolized a multi-ethnic South Africa , but one where living in dignity is shared across racial and class lines . Attacks against newcomers in South Africa are often reduced to attitudes of hate and resentment towards other black Africans . The national and international headlines use `` xenophobia '' as if any one word can convey the multifaceted crisis within which this phenomenon occurs . Labeling this turmoil as xenophobic fails to convey the conditions in which African migrants are scapegoated for the persistent legacy of apartheid in the post-liberation era . This word also does not tell us that extreme poverty now exceeds that experienced under apartheid . And it certainly does not account for how extreme inequality is now fully embraced and normalized by a new black elite joining the `` white-haves '' of yesteryear . The question of why foreign blacks are targeted and not foreign whites is also repeated ad nauseam , as if the majority of black citizens and African migrants share any common spaces and experiences with white South Africans or white foreigners . Contact between the majority of African migrant groups and native black South Africans mostly occurs in under-developed informal settlements and townships . In the interaction and competition between these groups in these spaces , we need to be wary of the simplistic treatment of South Africa 's ailment as xenophobia . Such labeling does not and can not explain the totality of the contact , competition and conflict between native poor black South Africans and foreign African entrepreneurs . Attention as to how such interactions occur in an environment where a vast portion of South Africa 's black majority experiences segregation , persistent and relative poverty , and high crime rates in post-apartheid South Africa is paramount . Neglecting access to social rights ,

Figure 8: Specialized NE head showing a non interpretable pattern.

Table 1: Metric scores for the decoder cross attention of model 1. #1 POS and #1 NE show the most attended POS tag or named entity for that attention head along with its ratio compared to the other tags. For each column, the three heads with the highest scores are boldfaced.

		POS-KL	NEP	NER-KL	#1 POS	#1 NE
Layer 0	Head 0	0.03 ± *0.02*	0.15 ± *0.08*	0.04 ± *0.05*	NOUN: 0.340	PER: 0.610
	Head 1	0.05 ± *0.03*	0.13 ± *0.07*	0.1 ± *0.09*	NOUN: 0.360	PER: 0.560
	Head 2	0.03 ± *0.02*	0.13 ± *0.08*	0.06 ± *0.09*	NOUN: 0.330	PER: 0.490
	Head 3	0.1 ± *0.04*	0.1 ± *0.06*	**0.21** ± *0.19*	NOUN: 0.240	PER: **0.760**
	Head 4	0.04 ± *0.03*	0.16 ± *0.09*	0.06 ± *0.05*	NOUN: 0.350	PER: 0.570
	Head 5	0.07 ± *0.03*	0.15 ± *0.08*	0.15 ± *0.13*	NOUN: 0.390	PER: 0.630
	Head 6	0.12 ± *0.05*	0.09 ± *0.05*	0.08 ± *0.08*	ADP: 0.240	PER: 0.430
	Head 7	0.09 ± *0.03*	0.16 ± *0.07*	0.1 ± *0.1*	NOUN: 0.350	PER: 0.520
Layer 1	Head 0	0.08 ± *0.04*	0.09 ± *0.06*	0.12 ± *0.11*	NOUN: 0.370	PER: 0.520
	Head 1	0.15 ± *0.06*	0.17 ± *0.09*	0.13 ± *0.11*	NOUN: 0.300	PER: 0.670
	Head 2	0.15 ± *0.05*	0.13 ± *0.08*	0.19 ± *0.15*	NOUN: 0.390	ORG: 0.420
	Head 3	0.07 ± *0.04*	0.15 ± *0.08*	0.2 ± *0.17*	NOUN: 0.350	PER: 0.720
	Head 4	**0.42** ± *0.14*	0.09 ± *0.05*	0.07 ± *0.07*	PUNC: 0.430	PER: 0.660
	Head 5	0.14 ± *0.06*	0.2 ± *0.09*	0.11 ± *0.12*	NOUN: 0.320	PER: 0.640
	Head 6	0.09 ± *0.06*	0.15 ± *0.07*	0.11 ± *0.1*	NOUN: 0.350	PER: 0.540
	Head 7	0.13 ± *0.04*	**0.27** ± *0.09*	0.15 ± *0.15*	NOUN: 0.380	ORG: 0.470
Layer 2	Head 0	0.15 ± *0.05*	**0.23** ± *0.09*	0.08 ± *0.1*	NOUN: 0.440	PER: 0.640
	Head 1	0.11 ± *0.06*	0.15 ± *0.08*	**0.21** ± *0.16*	NOUN: 0.230	PER: **0.780**
	Head 2	**0.25** ± *0.09*	**0.26** ± *0.13*	0.1 ± *0.1*	NOUN: **0.560**	PER: 0.610
	Head 3	0.09 ± *0.07*	0.12 ± *0.13*	0.13 ± *0.13*	NOUN: 0.290	PER: 0.680
	Head 4	0.18 ± *0.06*	0.18 ± *0.09*	0.11 ± *0.11*	NOUN: **0.480**	PER: **0.830**
	Head 5	0.14 ± *0.08*	0.16 ± *0.09*	0.1 ± *0.09*	NOUN: 0.390	PER: 0.590
	Head 6	0.06 ± *0.03*	0.12 ± *0.07*	**0.22** ± *0.19*	NOUN: 0.330	PER: 0.720
	Head 7	0.12 ± *0.07*	0.15 ± *0.11*	0.09 ± *0.1*	NOUN: 0.300	PER: 0.460
Layer 3	Head 0	0.07 ± *0.04*	0.15 ± *0.08*	0.2 ± *0.17*	NOUN: 0.350	PER: 0.690
	Head 1	0.17 ± *0.12*	0.09 ± *0.05*	0.11 ± *0.11*	PUNC: 0.230	PER: **0.760**
	Head 2	0.11 ± *0.06*	0.12 ± *0.09*	0.2 ± *0.18*	NOUN: 0.420	PER: 0.620
	Head 3	**0.19** ± *0.18*	0.2 ± *0.25*	0.16 ± *0.16*	NOUN: 0.350	ORG: 0.540
	Head 4	0.1 ± *0.06*	0.14 ± *0.09*	0.09 ± *0.08*	NOUN: 0.270	PER: 0.670
	Head 5	0.11 ± *0.06*	0.14 ± *0.06*	0.16 ± *0.14*	NOUN: 0.300	PER: 0.420
	Head 6	0.16 ± *0.07*	0.18 ± *0.1*	0.1 ± *0.1*	NOUN: **0.490**	PER: 0.680
	Head 7	0.07 ± *0.04*	0.13 ± *0.08*	0.19 ± *0.17*	NOUN: 0.360	PER: 0.750

Table 2: Metric scores for the decoder cross attention of model 2. #1 POS and #1 NE show the most attended POS tag or named entity for that attention head along with its ratio compared to the other tags. For each column, the three heads with highest scores are boldfaced.

		POS-KL	NEP	NER-KL	#1 POS	#1 NE
Layer 0	Head 0	0.04 ± *0.03*	0.14 ± *0.07*	0.04 ± *0.05*	NOUN: 0.320	PER: 0.480
	Head 1	0.05 ± *0.03*	0.18 ± *0.09*	0.06 ± *0.06*	NOUN: 0.370	PER: 0.580
	Head 2	0.06 ± *0.03*	0.13 ± *0.08*	**0.24** ± *0.2*	NOUN: 0.310	PER: 0.560
	Head 3	0.04 ± *0.02*	0.14 ± *0.06*	0.04 ± *0.04*	NOUN: 0.350	PER: 0.490
	Head 4	0.06 ± *0.03*	0.11 ± *0.07*	0.2 ± *0.17*	NOUN: 0.280	ORG: 0.490
	Head 5	0.05 ± *0.03*	0.19 ± *0.09*	0.08 ± *0.07*	NOUN: 0.380	PER: 0.580
	Head 6	0.17 ± *0.05*	0.14 ± *0.08*	0.06 ± *0.05*	NOUN: 0.290	PER: 0.690
	Head 7	0.1 ± *0.04*	0.17 ± *0.07*	0.06 ± *0.06*	NOUN: **0.410**	PER: 0.520
Layer 1	Head 0	0.18 ± *0.07*	0.16 ± *0.07*	0.09 ± *0.08*	PUNC: 0.290	PER: 0.560
	Head 1	0.14 ± *0.07*	0.16 ± *0.08*	0.08 ± *0.07*	NOUN: 0.370	PER: 0.620
	Head 2	0.09 ± *0.04*	0.17 ± *0.09*	0.13 ± *0.12*	NOUN: 0.400	PER: 0.600
	Head 3	0.1 ± *0.05*	0.19 ± *0.07*	0.11 ± *0.1*	NOUN: 0.310	PER: 0.440
	Head 4	0.06 ± *0.03*	0.11 ± *0.06*	0.06 ± *0.06*	NOUN: 0.360	PER: 0.620
	Head 5	0.17 ± *0.07*	0.17 ± *0.1*	0.1 ± *0.1*	NOUN: 0.370	PER: 0.710
	Head 6	**0.19** ± *0.08*	0.13 ± *0.06*	0.1 ± *0.1*	VERB: 0.340	PER: 0.670
	Head 7	**0.22** ± *0.1*	0.17 ± *0.08*	0.1 ± *0.1*	VERB: 0.320	PER: **0.740**
Layer 2	Head 0	0.08 ± *0.03*	0.18 ± *0.09*	0.1 ± *0.09*	NOUN: 0.400	ORG: 0.500
	Head 1	0.05 ± *0.03*	0.1 ± *0.05*	0.07 ± *0.07*	NOUN: 0.300	PER: 0.670
	Head 2	0.13 ± *0.07*	0.13 ± *0.08*	0.08 ± *0.08*	NOUN: 0.310	PER: **0.830**
	Head 3	0.05 ± *0.02*	0.16 ± *0.08*	0.08 ± *0.08*	NOUN: 0.290	PER: 0.550
	Head 4	0.06 ± *0.03*	0.12 ± *0.07*	**0.21** ± *0.18*	NOUN: 0.300	PER: 0.520
	Head 5	0.14 ± *0.05*	0.16 ± *0.08*	0.08 ± *0.08*	NOUN: 0.390	PER: 0.610
	Head 6	**0.24** ± *0.09*	**0.23** ± *0.12*	0.07 ± *0.08*	NOUN: 0.430	PER: 0.690
	Head 7	0.09 ± *0.04*	0.1 ± *0.06*	**0.21** ± *0.19*	NOUN: 0.340	PER: **0.760**
Layer 3	Head 0	0.08 ± *0.05*	0.16 ± *0.09*	0.2 ± *0.17*	NOUN: 0.360	PER: 0.470
	Head 1	0.11 ± *0.05*	**0.21** ± *0.09*	0.12 ± *0.12*	NOUN: **0.410**	PER: 0.680
	Head 2	0.12 ± *0.12*	**0.24** ± *0.18*	0.11 ± *0.12*	NOUN: **0.420**	PER: 0.490
	Head 3	0.15 ± *0.07*	0.17 ± *0.09*	0.16 ± *0.15*	NOUN: 0.400	PER: 0.490
	Head 4	0.07 ± *0.03*	0.11 ± *0.06*	**0.21** ± *0.18*	NOUN: 0.370	PER: 0.570
	Head 5	0.08 ± *0.09*	0.12 ± *0.08*	0.1 ± *0.11*	NOUN: 0.290	PER: 0.510
	Head 6	0.1 ± *0.05*	0.13 ± *0.08*	0.11 ± *0.11*	NOUN: 0.360	PER: 0.700
	Head 7	0.07 ± *0.03*	0.12 ± *0.07*	0.13 ± *0.12*	NOUN: 0.330	PER: 0.730