A Multiscale Visualization of Attention in the Transformer Model-replicated study

Analysis by : Mohammed Boudjemai, Alex Liu, Said Sadeg

March 7, 2021

1. Introduction

Natural Language Processing (NLP) is one of the fastest growing fields of artificial intelligence. It is centered on the understanding of human language, represented by textual data, by computers. Question Answer (QA) is a NLP subfield in which an algorithm receives a short text (a context) and is trained to answer natural questions relating to that text by highlighting the most relevant passage to answer the question in the context text. This approach can be generalized to Open-Domain QA (or large Close Domain QA), with the added difficulty that the context text is not given and instead must be retrieved from a much larger set of documents by the algorithm. Although the field of NLP research is extremely active, most of the effort is focused around the English language.

However, it would be wrong to say that no effort is made in other languages. Recently, state-of-the-art NLP architectures have been adapted to the French language (CamemBERT and FlauBERT) and have paved the way for research on a variety of tasks that were not possible before).

In this work, we present the work of Jesse Vig (2019) who introduced a tool for visualizing attention in the Transformer at multiple scales. he demonstrated the tool on GPT-2 and BERT, and he presented three use cases. However, Jesse introduced a high-level model view, which visualizes all layers and heads of attention in a single interface, and a low-level neuron view, which shows how individual neurons interact to produce attention. He also adapt the tool from the original encoder-decoder implementation to the GPT-2 decoder-only model and then the BERT encoder-only model.

2. State of the art

The recent popularity of intelligent assistants has increased interest in question-and-answer systems (QASs) which have become a central element of ?Human-Machine? exchanges since they allow users to have direct answers to their questions by natural language using their own terminology without having to go through a long list of documents to find the appropriate answers.

BERT (Bidirectional Encoder Representations from Transformers), is designed to pre-train deep two-way representations from unlabeled text by jointly conditioning left and right context in all layers. As a result, the pre-trained BERT model can be refined with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as answering questions and language inference, without substantial modifications to the task-specific architecture.

BERT is conceptually simple and empirically powerful. It achieves new cutting-edge results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (absolute improvement of 7.7%), MultiNLI accuracy to 86.7% (absolute improvement of 4, 6%), SQuAD v1.1 question answering Test F1 to 93.2 (absolute improvement of 1.5 points) and SQuAD v2.0 Test F1 to 83.1 (absolute improvement of 5.1 points) (Devlin et al.,2018). Several attempts have been made to develop generative antagonistic networks (RAG) or generative neural recurrent networks (RRNG) for textual data such as TextGAN [Zhang et al, 2017]. There is also a platform, called TexyGen, to compare the different approaches [Zhu et al, 2018], [Texygen, 2018].

A deep learning expert confirms, in theory, that a network is capable of learning if it has enough relevant data. But currently, the generation of sentences from recurrent networks is slow to produce meaningful sentences, even considering the progress recently announced by OpenAI with its generative model GPT-2 (Generative Pre-training Transformers) [Radford et al, 2019].

Although several models developed for other languages ??have been published (ELMo 1 models for Japanese, Portuguese, German and Basque; BERT models for Simplified and Classical Chinese (Devlin et al., 2018) or for the German (Chan et al., 2019)), the differential in the size of their pre-training data did not allow the emergence of studies comparing them to the original model. However, multilingual models based on the concatenation of large datasets (mainly based on Wikipedia) have appeared (Devlin et al., 2018; Conneau et al., 2019) and have enabled notable advances through transfer learning (Pires et al., 2019). However, it is only very recently that large-scale monolingual models have been developed (Martin et al., 2019; Le et al., 2019; Virtanen et al., 2019; Delobelle et al., 2020) and have allowed confirming the interest of monolingual models on other languages.

With regard to the French language, Le et al. (2019) have shown on various tasks that their model, FlauBERT, offered a panel of performances equivalent to those of CamemBERT (Martin et al., 2019), emphasizing the complementarity of the two models on parsing tasks.

CAMEMBERT is based on ROBERTA (Liu et al., 2019), an evolution of BERT (Devlin et al., 2019) on several levels, in particular by the use of the masked language model as the only pre-training objective. In addition to the original CAMEMBERTBASE model driven with 12 layers, 768 hidden dimensions and 12 attention heads, or 110M parameters.

One challenge for visualizing attention in the Transformer is that it uses a multi-layer, multihead attention mechanism, which produces different attention patterns for each layer and head. BERT-Large, for example, which has 24 layers and 16 heads, generates 24 * 16 = 384 unique attention structures for each input. Jones (2017) designed a visualization tool specifically for multihead attention, which visualizes attention over multiple heads in a layer by superimposing their attention patterns (Vaswani et al., 2017, 2018).

Various tools have been developed to visualize attention in NLP models, ranging from attention-matrix heatmaps (Bahdanau et al., 2015; Rush et al., 2015; Rocktaschel et al. , 2016) to bipartite graph representations (Liu et al., 2018; Lee et al., 2017; Strobelt et al., 2018).

Rajaswa Patil et al., 2020, described an approach to model causal reasoning in natural language by detecting counterfactuals in text using multi-head self-attention weights. they used pre-trained transformer models to extract contextual embeddings and self-attention weights from the text. they showed the use of convolutional layers to extract task-specific characteristics from these self-attention weights.

3. Summary of the replicated study

To visualizing attention for individual inputs to the model, we also analyze attention in aggregate the following research questions:

3.1. Transformer Architecture

For the Stacked Decoder, GPT-2 is a stacked decoder Transformer, which inputs a sequence of tokens and applies position and token embeddings followed by several decoder layers. Each layer applies multi-head self-attention (see below) in combination with a feedforward network, layer normalization, and residual connections. The GPT-2 small model has 12 layers and 12 heads.

For Given an input x, the self attention mechanism assigns to each token xi a set of attention weights over the tokens in the input:

$Attn(xi) = (\alpha i1(x), \alpha i2(x), ..., \alpha ij(x))$

where $\alpha ij(x)$ is the attention that xi pays to xj. The weights are positive and sum to one. Attention in GPT-2 is right-to-left, so it is is defined only for $j \leq i$. In the multi-layer, multi-head setting, ? is specific to a layer and head.

The attention weights $\alpha ij(x)$ are computed from the scaled dot-product of the query vector of $xi$and the key vector of xj, followed by a softmax operation. The attention weights are then used to produce a weighted sum of value vectors:

$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$

using query matrix Q, key matrix K, and value matrix V , where dk is the dimension of K. In a multi-head setting, the queries, keys, and values are linearly projected h times, and the attention operation is performed in parallel for each representation, with the results concatenated.

3.2. Visualizing Individual Inputs

For the model view visualizes attention across all of the model?s layers and heads for a particular input. Attention heads are presented in tabular form, with rows representing layers and columns representing heads. Each head is shown in a thumbnail form that conveys the coarse shape of the attention pattern, following the small multiples design pattern.

For the neuron view visualizes how individual neurons interact to produce attention. This view displays the queries and keys for each token, and demonstrates how attention is computed from the scaled dot product of these vectors. The element-wise product shows how specific neurons influence the dot product and hence attention, where the attention distribution at position i in a sequence x is defined as follows:

$\alpha_i = softmax(\frac{q_i k_1}{\sqrt{d}}, \frac{q_i k_2}{\sqrt{d}}, ..., \frac{q_i k_N}{\sqrt{d}})$

where q_i is the query vector at position i, k_j is the key vector at position j, and d is the dimension of k and q. N=i for GPT-2 and N=len(x) for BERT. All values are specific to a particular layer / head.

The columns in the visualization are defined as follows:

- Query q: The query vector of the selected token that is paying attention.?
- Key k: The key vector of each token receiving attention.?
- q*k (element-wise): The element-wise product of the query vector and each key vector.
- This shows how individual neurons contribute to the dot product (sum of element wise product) and hence attention.?
- q.k: The dot product of the selected token?s query vector and each key vector.?
- Softmax: The softmax of the scaled dotproduct from previous column. This is the attention score. ?

?

4. Results

4.1. Attention-head view

In this section we describe the different simulations performed in this study which includes three views: a seen attention head, a model view and a neuron view. in what follows, we will describe these points of view and demonstrate them on the GPT-2 and BERT models.

First with Bert, In this experiment, we tested the phrases ?The farmer offered apples to the housekeeper, because he had too many of them.? and "The farmer asked the designer what she was working on" and we were able to see links between the pronouns with their antecedent, in particular between the pronoun ”he” and the antecedent ”farmer” in layer 5 of the network and we have also a link between the pronoun ”she” and the antecedent ”housekeeper” in the same layer, Figure.1.

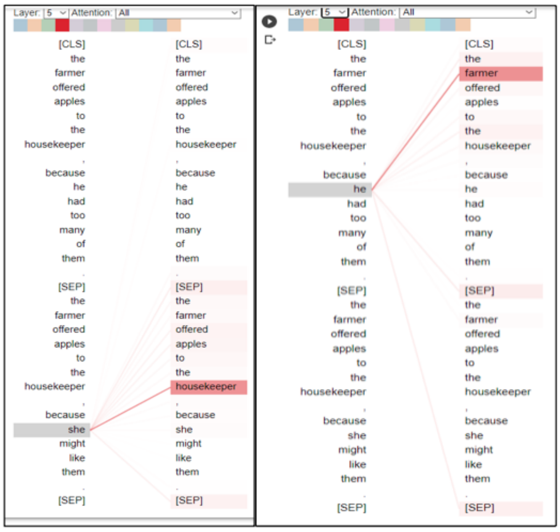

Figure 1: Representation of different layers / attention heads (The link with the first name and the antecedent with bert)

With GPT2 model, we tested with the sentence ?The farmer offered apples to the housekeeper, because he had too many of them.? and we have seen a link between the pronoun ”he” and the previous ”farmer” in layer 5 of the network, Figure.2.

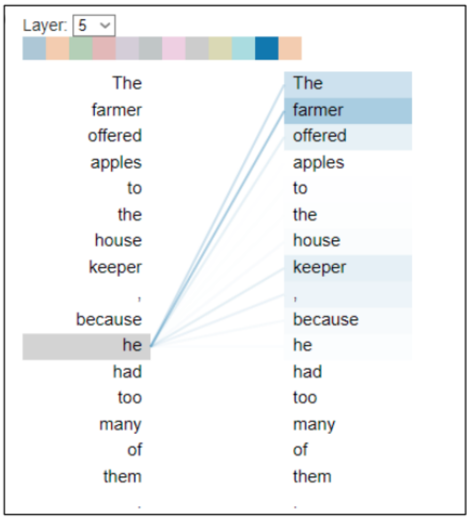

Figure 2: The link with the first name and the past with gpt2

4.2. Neuron view

With GPT2 By visualizing closely a neuron taken in layer 5, we notice that the link between the pronoun and its antecedent is well reflected at the neuron level, indeed we can clearly see that the highest value of the products kxq corresponds to the antecedent ”farmer ”Of” he ”, Figure.3, Figure.4.

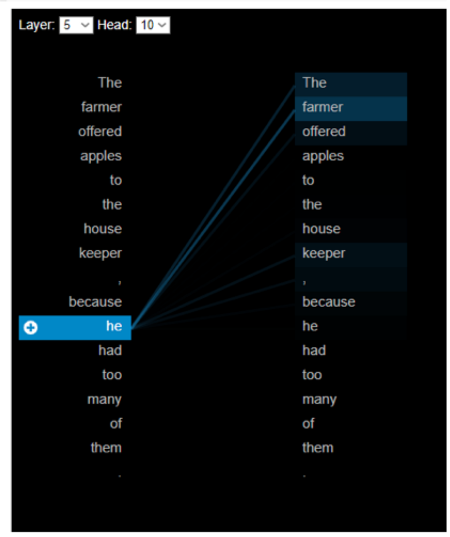

Figure 3: Neuron view with gpt2

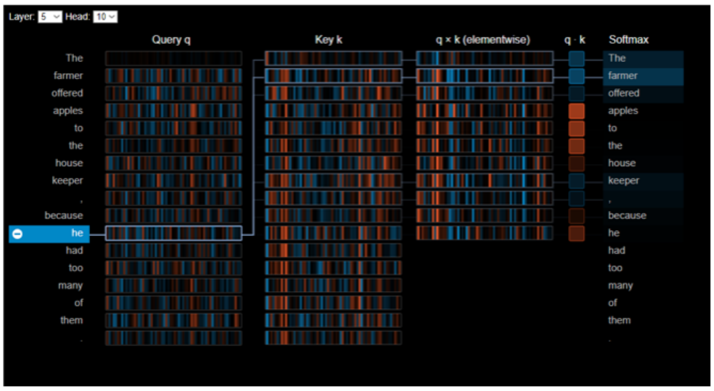

Figure 4: Neuron view details gpt2

4.3. Model View

The Model gives us a general overview of all the layers with the different attention heads, we can also observe the variation of the links between the different words of each sentence. This allows us to easily identify the layer in which we have a link between the pronoun and its antecedent, Figure 5.

5. Issues and Discussion

By executing the code provided with the paper ”Vig, J. (2019). A multiscale visualization of attention in the transformer model. ”, With sentences chosen in github data.
we can see that the results correspond well to the result presented in the paper, the results can be reproduced without error. However, when the sentences are long we find that the execution becomes slower and in some cases can lead to a display problem.

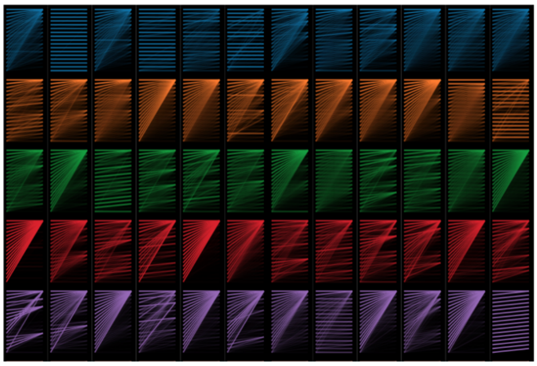

Figure 5: Representation of Model View

By analyzing the links between the pronouns and their antecedent in the two models (Bert and GPT2), we see that these links are both in the same layer, which leads us to say that the two models have a similar behavior, and therefore allows to have confidence in the results of these models.

6. Conclusion

In this replicated study, we have presented a tool for visualizing attention in the Transformer at multiple scales. We demonstrated the tool on GPT-2 and BERT, and presented several use cases.

We have found that many attention heads specialize in particular part of speech tags and that different tags are targeted at different layer depths. We also found that the deeper layers capture the more distant relationships and that attention aligns most strongly with the dependency relationships in the middle layers where the attention distance is smallest. Our qualitative analysis revealed that the structure of attention is closely related to the training objective; for GPT-2, which was trained using left-to-right language modeling, attention was often focused on the words most relevant to predicting the next token in the sequence.

We believe that the interpretation of an attention-based model is complementary to linguistic probing approaches. While linguistic probing precisely quantifies the amount of information encoded in various components of the model, it requires the training and evaluation of a probing classifier. Attention analysis is a simpler process that also produces human-interpretable descriptions of the model's behavior.

The results of our analyzes were often consistent with those of the survey approaches.

7. Acknowledgements

This work was supported by Paris Descartes university. We would especially like to thank our advisor, Prof.Nicolas Ballier for his proposed paper, commitment, and unlimited patience. We are also grateful for the help and wonderful device given by all colleagues for devoting their time to read our work and for their valuable suggestions.

7.1. References

1. CHAN B., MOLLER T., PIETSCH M., SONI T. YEUNG C. M. (2019). German bert. https://deepset.ai/german-bert.

2. CHARNIAK E. (2019). Introduction to deep learning. The MIT Press.

3. CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G., GUZMAN F.,GRAVE E., OTT M., ZETTLEMOYER L. STOYANOV V. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint : 1911.02116.

4. CONNEAU A., RINOTT R., LAMPLE G., WILLIAMS A., BOWMAN S. R., SCHWENK H., STOYANOV V. (2018). XNLI : evaluating cross-lingual sentence representations. In E. RILOFF, D.

5. CHIANG, J. HOCKENMAIER J. TSUJII, eds., Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, p. 2475?2485 : Association for Computational Linguistics.

6. DAI A. M. LE Q. V. (2015). Semi-supervised sequence learning. In Advances in Neural , Information Processing Systems 28 : Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, p. 3079?3087.

7. DELOBELLE P., WINTERS T. BERENDT B. (2020). RobBERT : a Dutch RoBERTa-based Language Model. arXiv preprint : 2001.06286. DEVLIN J., CHANG M., LEE K. TOUTANOVA K. (2018). Multilingual bert. https://github.com/google-research/bert/blob/master/multilingual.md.

8. DEVLIN J., CHANG M., LEE K. TOUTANOVA K. (2019). BERT : pre-training of deep bidirectional transformers for language understanding. In J. BURSTEIN, C. DORAN T. SOLORIO, eds.,?Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), p. 4171?4186 : Association for Computational?Linguistics.

9. DOZAT T. MANNING C. D. (2017). Deep biaffine attention for neural dependency parsing. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings : OpenReview.net.

10. GRAVE E., BOJANOWSKI P., GUPTA P., JOULIN A. MIKOLOV T. (2018). Learning word vectors for 157 languages. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018. : European Language Resources Association (ELRA).

11. GRAVE E., MIKOLOV T., JOULIN A. BOJANOWSKI P. (2017). Bag of tricks for efficient

text classification. In M. LAPATA, P. BLUNSOM A. KOLLER, eds., Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2 : Short Papers, p. 427-431 : Association for?Computational Linguistics.

12. HOWARD J. RUDER S. (2018). Universal language model fine-tuning for text classification. In I. GUREVYCH Y. MIYAO, eds., Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1 : Long Papers, p. 328-339 : Association for Computational Linguistics. doi : 10.18653/v1/P18-1031.

13. JAWAHAR G., SAGOT B., SEDDAH D., UNICOMB S., INIGUEZ G., KARSAI M., LEO Y., KARSAI M., SARRAUTE C., FLEURY . et al. (2019). What does bert learn about the structure of language ? In 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy. Jesse Vig, A Multiscale Visualization of Attention in the Transformer Model, arXiv:1906.05714v1 [cs.HC] 12 Jun 2019

14. JOULIN A., GRAVE E., BOJANOWSKI P., DOUZE M., JEGOU H. MIKOLOV T. (2016). Fasttext.zip : Compressing text classification models. arXiv preprint : 1612.03651.

15. KINGMA D. P. BA J. (2014). Adam : A method for stochastic optimization. arXiv preprint : 1412.6980.

16. LAN Z., CHEN M., GOODMAN S., GIMPEL K., SHARMA P. SORICUT R. (2019). ALBERT : Alite BERT for self-supervised learning of language representations. arXiv preprint : 1909.11942.

17. LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBE B., BESACIER L. SCHWAB D. (2019). Flaubert : Unsupervised language model pre-training for french. arXiv preprint : 1912.05372.

18. LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. STOYANOV V. (2019). Roberta : A robustly optimized BERT pretraining approach. arXiv preprint : 1907.11692.

19. MARTIN L., MULLER B., ORTIZ SUAREZ P. J., DUPONT Y., ROMARY L., VILLEMONTE DE LA CLERGERIE ., SEDDAH D. SAGOT B. (2019). CamemBERT : a Tasty French Language Model.?arXiv preprint : 1911.03894.

20. MIKOLOV T., GRAVE E., BOJANOWSKI P., PUHRSCH C. JOULIN A. (2018). Advances in pre-training distributed word representations. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.

21. MIKOLOV T., SUTSKEVER I., CHEN K., CORRADO G. S. DEAN J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. BURGES, L. BOTTOU, Z. GHAHRAMANI K. Q. WEINBERGER, eds., Advances in Neural Information Processing?Systems 26 : 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States., p. 3111-3119

22. Rajaswa Patil et al., CNRL at SemEval-2020 Task 5: Modelling Causal Reasoning in Language with Multi-Head Self-Attention Weights based Counterfactual Detection, arXiv:2006.00609v1 [cs.CL] 31 May 2020