Replication of the paper : A Multiscale Visualization of Attention in the Transformer Model

BOUAYAD Ghita, KERAGHEL Imed, BAMOUH Mohamed, HEBROUNE Oussama
Paris University, 45 Rue des Saints-Pères, 75006 Paris
bouayad.ghita0@gmail.com, keraghel.imed19@gmail.com,
bamouh42@gmail.com, oussamaheb@gmail.com

Abstract

Recent work on the natural language processing field allowed the emergence of new powerful language models that excel in encoding large documents, even taking into consideration their lexical nuances, and decoding them in the form of coherent essays. In order to highlight those advancements, we read and replicated the results of the paper "A Multiscale Visualization of Attention in the Transformer Model" (Vig, 2019), where the research team at the Palo Alto Research Center visualized the attention mechanisms of the transformer model at various scales and stages.

Keywords: BERT, GPT-2, Visualization

1. Introduction

Since 2017, Attention models (Vaswani et al., 2007), and especially its variants, such as BERT (Devlin et al., 2018) and OpenAI GPT-2 model (Radford et al., 2019), provide state of the art result on tasks relative to the natural language processing field. In this report, we delved into attention mechanisms, and visualized the self-attention outputs at different layers of the encoder in the BERT model as well as the decoder stack in the GPT-2 model. We mainly focused on detecting model bias and pronoun matching in medium sized sentences using the results of the visualization tool.

2. State of the Art

2.1. Recurrent Neural Networks

For a long time, the Recurrent Neural Network architecture (Rumelhart et al., 1985) provided faithful contextual representation of sequential data, which could be then applied to a variety of tasks, such as text classification using Many-to-One models, or text generation using Many-to-Many models. Unfortunately, those models showed poor performance when applied to medium to long sequence of texts.

2.2. LSTM and GRU

In order to remedy to this, GRU (Cho et al., 2014) and LSTM (Hochreiter and Schmidhuber, 1997) were used to supply the model with "memory cells" so that it could retain context at the beginning of the sentence over long chunks of text. But even those models performed poorly on multi-document text classification and human-readable, coherent text generation, in addition to being slow to train, frequently subject to vanishing/exploding gradients, and not suitable for transfer learning.

2.3. Attention Models

Enter the transformer model, presented in the "Attention is All you Need" paper (Vaswani et al., 2007), which resolves all the aforementioned problems by introducing an architecture split in two main components: **the Encoder**, which encodes the contextual representation of large chunks of text using self-attention heads, and **the Decoder**, which relies on masked self-attention, encoder-decoder attention and a softmax function to output text. The encoder supports parallelization for faster training whereas the Decoder is more sequential.

2.4. BERT and GPT

2.4.1. BERT

The Bidirectional Encoder Representations from Transformers model (Devlin et al., 2018) provides state of the art results on a wide variety of natural language processing tasks. It relies on the Encoder part of the Transformer's model architecture and expands on it by stacking multiple encoder layers, and training them on two tasks: Masked Language Model and Next Sentence Prediction. The resulting context from running the BERT Model on a sentence or a document can be fed to a custom neural net in order to adapt to the current task, be it text classification, translation or named entity recognition.

2.4.2. GPT

Even though the paper we're replicating uses the OpenAI GPT-2 model, a more recent model, the GPT-3 model (Brown et al., 2020), provides even more impressive results in tasks involving text generation, by printing out syntactically coherent text. It also relies on the Transformer architecture, but unlike BERT, GPT-2 and GPT-3 use Decoder layers as building blocks for its language model.

3. Summary of the replicated study

The paper provides an open-source, interactive visualization tool that illustrates the attention mechanisms of the BERT model as well as the GPT-2 model at various stages, heads and layers. Two views are available:

- A high-level model view, which visualizes all of the layers and attention heads in a single interface,
- A low-level neuron view, which shows how individual neurons interact to produce attention.

Studying the two views using the tool allows us to efficiently highlight lexical, syntactic and semantic patterns detected by the two models on various examples, and detect potential relevant attention heads in the self-attention process of the BERT and GPT-2 models.
In particular, one of the use cases for attention visualization is to detect potential model bias.
In the paper, gender bias was uncovered in the GPT-2 model, indeed, when fed sentences relative to the medical field, the model associated male pronouns to doctors and female pronouns to nurses.
Another use case is to run diagnostic checks on the learning process, by studying the values and the evolution of the weights in each individual neuron. This allows us to link neuron activity to model behavior.

4. Results and Discussion

4.1. GPT-2 model

The GPT model recognizes gender biases for most of *en_pro* dataset's sentences as shown in the pictures below (cf. figure 1). However, for some sentences, it fails to map the pronoun with the referrant's gender (cf. figure 2).

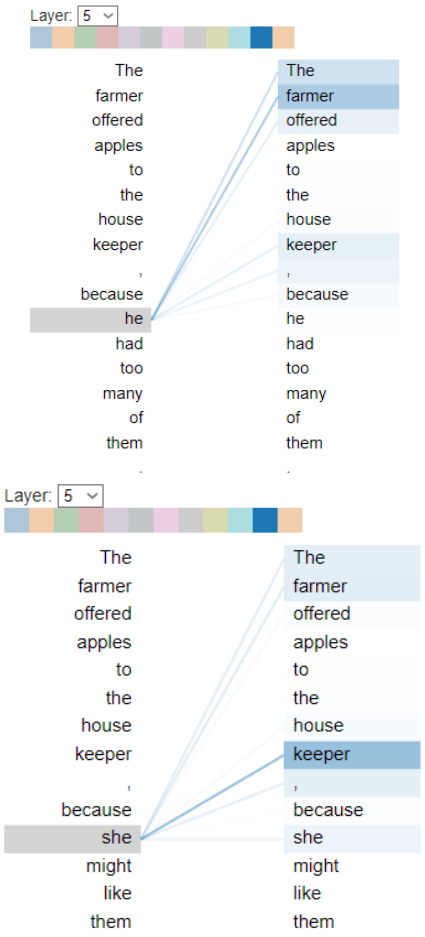

Figure 1: Self-attention results in one of the attention heads of the fifth layer of the GPT-2 model

Here, the pronouns match their respective subjects, gender bias isn't explicitly shown, but it doesn't means that it isn't applied, since the *farmer* profession may be perceived by the model as strongly attributed to the "he" pronoun. We need other examples to highlight more obvious gender bias.

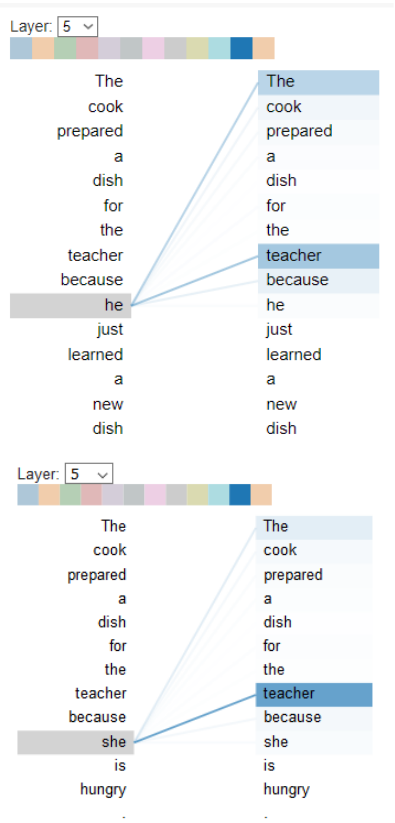

Figure 2: Pronoun mismatched to the associated noun in the GPT-2 model

This figure clearly display of the GPT-2 model showing gender bias. The model inverted the pronouns and the professions. It associated the male pronoun "he" with the *teacher* profession, and the "she" pronoun with the *cook* profession.

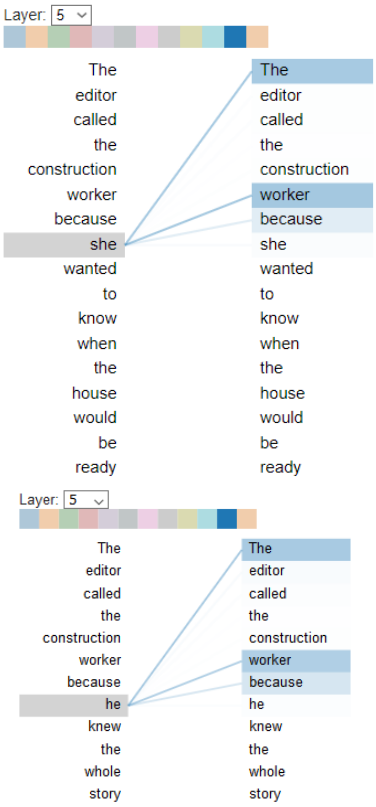

Figure 3: Pronoun mismatched to the associated noun in the GPT-2 model

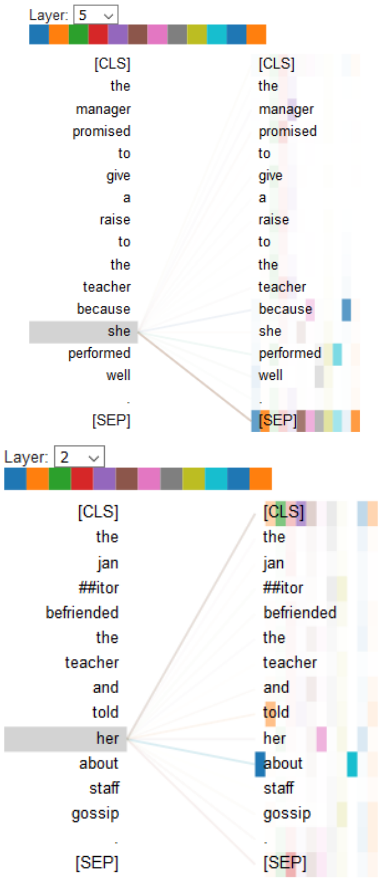

Figure 4: Pronoun mismatched in the BERT model for every layer and attention head

This is an interesting example. Both the gender pronouns are strongly associated with the *worker* profession, regardless of their position. This is probably another kind of bias that is at work, one that gives more weight to general professions, such as *worker, keeper or employee*, than specific ones.

4.2. BERT model

As for the BERT model, it couldn't detect neither the lexical patterns in the sentences nor the bias gender in any of attention heads.

The BERT model's self-attention heatmaps seem to point to the special tokens [SEP] and [CLS]. There doesn't seem to be any kind of pronoun matching, and by extension, any kind of eventual model bias.

5. Encountered Issues

During this study we encountered many obstacles. First of all, since Bertviz returns HTML and JavaScript objects, we couldn't use Docker to visualize the output, as they are only adapted to a browser. Another issue, the RAM was often crashing due to the big amount of dataset and the costly computation of long sequences. Moreover, in order to evaluate models we had to inspect manually each sentence's visualization, which is time consuming and inconvenient, as we didn't have time to implement an explicit metric to quantify the presence of gender bias in our dataset.

6. Conclusion

The results we obtained seem to match the results seen replicating the original paper. The GPT-2 model displays better capacity to detect lexical patterns in the given examples than the BERT model. In the original paper, only GPT-2 was used for gender bias detection. In our study, we made a comparative analysis of both BERT and GPT-2's capacity to show gender bias. We found that only GPT-2 highlighted obvious gender bias when applied to sentences from the mt-gender dataset. As for the BERT model, we

haven't been able to find any kind of gender bias when running it on multiple examples from the dataset.

7. Acknowledgements

This work would not have been possible without the paper and code of the original article (Vig, 2019). Everyone in the student team contributed to the realisation of the study. The code was written and documented by Ghita Bouayad and Imed Keraghel. The reading of the original paper as well as the writing of the report was mainly done by Mohamed Bamouh and Oussama Hebroune.

8. Bibliographical References

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers forlanguage understanding.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation.

Vaswani, A., Noam Shazeer, N. P., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2007). Attention is all you need.

Vig, J. (2019). A multiscale visualization of attention in the transformer model.

9. Dataset used

`https://github.com/gabrielStanovsky/mt_gender/blob/master/data/aggregates/en_pro.txt`