Plausibility Processing in Transformers: Where is this ability coming from?

Anonymous ACL submission

Abstract

Transformers are found to process semantic knowledge in a human-like way. However, it has not been examined *where* and *how* semantic knowledge is processed inside the model. This paper aims to deepen understanding of how Transformers preserve and process semantic knowledge, focusing on semantic knowledge of plausible relations between nouns and verbs. In particular, I investigate how knowledge of semantic plausibility is localized in Transformer models and how such localized components make causal contributions to Transformers' plausibility processing ability. A set of experiments showed that i) Transformers have attention heads that detect plausible relations between nouns and verbs, and that ii) they collectively contribute to the Transformer's ability to process plausibility, though each attention head makes different amount of contribution.

1 Introduction

Transformers are attention-based neural network models (Vaswani et al., 2017), and they have brought breakthroughs in the field of Natural Language Processing achieving state-of-the-art performance in diverse downstream tasks such as machine translation, sentiment analysis, and text summarization, to name a few. Such great performance is mainly attributed to Transformers' ability to build dependencies even between long-distant words which attention heads are developed for (Merkx and Frank, 2020). To be specific, unlike previous neural network language models (e.g., Simple Neural Networks or Recurrent Neural Networks) that have issues retaining linguistic information coming from distant tokens, attention heads in Transformers enable to represent the meaning of tokens by integrating their contextual information without losing information from distant tokens (Bahdanau et al., 2014).

Provided that Transformer language models consist of multiple attention heads that serve different roles, previous studies examined functions that individual attention heads serve and how language processing work is divvied up inside Transformers (Clark et al., 2019; Voita et al., 2019; Vig, 2019; Jo and Myaeng, 2020). However, to the best of my knowledge, previous studies mostly focused on finding attention heads specialized for processing linguistic knowledge intrinsic to language systems, and little attention has been paid to semantic plausibility processing ability, which requires much of world knowledge going beyond linguistic knowledge. Consequently, we do not have yet clear answers to the origin of Transformers' general ability to process semantic plausibility in a human-like way, which has been observed in a number of studies (Bhatia et al., 2019; Misra et al., 2020, 2021; Han et al., 2022; Bhatia and Richie, 2022; Ralethe and Buys, 2022; Ettinger, 2020; Peng et al., 2022).

In this regard, the present study aims to fill the gap in our knowledge of Transformers' semantic processing by answering the following questions: (i) Are there attention heads specialized for processing semantic plausibility? and (ii) Do these heads actually generate causal effects on Transformers' ability to process semantic knowledge? Among many different types of linguistic knowledge that relates to semantic plausibility, this study particularly focuses on semantic knowledge that determines whether a noun and a verb are in a plausible relation (i.e., whether the semantic properties that a noun has match the ones that a verb has. See Section 3.1 for examples). To answer the first question, I examine which attention heads can detect plausible nouns over implausible ones in the most accurate and sensitive way. The second question is answered by investigating how Transformers' sensitivity to semantic plausibility changes as plausibility-processing heads are pruned.

A set of experiments uncover that Transformers have attention heads specialized for processing semantic plausibility, which are relatively diffusely

distributed from the bottom layers to the top layer. In addition, those attention heads are found to exert causal effects on Transformers' semantic plausibility processing ability, since Transformers' plausibility processing ability almost disappeared when the plausibility-processing attention heads are pruned.

In what follows, I will provide background that relates to questions I am addressing in this paper. In Section 3, I will conduct an experiment to find attention heads specialized for processing semantic plausibility knowledge and examine how they are distributed inside the model. In Section 4, I will examine the causal effects of the attention heads specialized for plausibility processing on Transformers' sensitivity to plausibility by examining changes in plausibility-sensitivity patterns of GPT2 models with different sets of attention heads. In section 5, I will summarize the results and discuss limitations of the study.

2 Background

2.1 How do attention heads work?

Attention heads are core components of Transformer which have led to great improvement in neural network language models by enabling context-dependent word representation with minimizing information loss from distant tokens.

In an attention head, each input token is multiplied by weight matrices (W_k,W_q, W_v) to construct key, query, and value vectors of the token. The query vector of a token, is then compared with key vectors of other input tokens by computing cosine similarities. This similarity score is what is called '*attention*' and it determines how much information should be extracted from a certain input token to build a contextual representation of the token being processed.

2.2 What roles do attention heads serve?

There have been a lot of studies that attempted to explain the language processing mechanism in Transformers with analyzing functions distinct attention heads serve (Voita et al., 2019; Vig, 2019; Clark et al., 2019; Jo and Myaeng, 2020). Specifically, Voita et al. (2019) found attention heads specialized for a position, syntactic relation, rare words detection; Vig (2019) found attention heads specialized in part-of-speech and syntactic dependency relation; Clark et al. (2019) found attention heads specialized in coreference resolution; and Jo and Myaeng (2020) examined how linguistic properties at sentence level, such as length of sentence, depth of syntactic tress, and so on, are processed in attention heads.

Despite numerous attempts in examining the roles of attention heads, the focus has been mostly on linguistic knowledge intrinsic to language systems which does not require much of world knowledge related to semantic processing. Thus, in order to account for where Transformer's ability to process semantic knowledge that relates to world knowledge, it needs to be closely examined how Transformers preserve and process such knowledge that can facilitate sentence processing.

2.3 How do we know whether attention heads are specialized for certain linguistic knowledge?

In previous studies, attention heads are considered to be specialized for a certain type of linguistic knowledge if attention distribution patterns in the attention heads are consistent with the linguistic knowledge (Voita et al., 2019; Vig and Belinkov, 2019). However, such regional analysis does not explain how much contribution attention heads make to Transformers' ability to process linguistic knowledge because such information from the specialized attention heads may fade away along with the information flows - from bottom layers to top layers, eventually making little contribution to Transformers' ability to process the linguistic knowledge.

Thus, in order to better understand how a set of components in language models are contributing to processing certain linguistic information, it is indispensable to analyze the causal effects that the attention heads can make on Transformer's ability to process linguistic information (Belinkov and Glass, 2019; Vig et al., 2020). In this sense, this paper will not only examine which attention heads can form attention distributions that are consistent with semantic plausibility knowledge, but also examine how much influence the attention heads can exert on Transformers' general ability to process plausibility.

This is a novel approach to investigating the role of individual attention heads because previous studies (Voita et al., 2019; Jo and Myaeng, 2020) showed how pruning important attention heads influences models' performance in downstream tasks without studying how the general linguistic ability is affected by the removal of attention heads specialized for that specific linguistic knowledge.

3 Plausibility Processors in Transformer

This experiment aims to find attention heads that are specialized for semantic plausibility processing. In particular, it will be examined whether there are attention heads that can form attention distribution patterns consistent with semantic knowledge of plausible relations between nouns and verbs, irrespective of syntactic dependency relation.

3.1 Data

Cunnings and Sturt (2018) examined how human sentence processing is affected by plausibility of true syntactic dependents of verbs and distractors. There are 32 sets of sentences with varying plausibility between correct dependents and verbs and plausibility between distractors and verbs[1].

For instance, in (a), the verb *shattered* forms a dependency with *plate*, and there is a distractor, *cup*, that could be erroneously considered as a dependent of *shattered* because of functional similarities with correct dependent. In (a), correct dependent (*plate*) and distractor (*cup*) have a feature [+shatterable], which makes both of them build plausible relation with the verb (*shattered*). In (d), however, the grammatically correct dependent *letter* and the distractor *tie* do not have a feature [+shatterable], and thus both of them are implausible dependent of the verb (*shattered*). By manipulating the plausibility of the correct dependent and the distractor, each set of sentences is with four conditions as shown in (a)-(d).

(a) *plausible - plausible*
Sue remembered the **plate** that the butler with the cup accidentally **shattered** ...

(b) *plausible - implausible*
Sue remembered the **plate** that the butler with the tie accidentally **shattered** ...

(c) *implausible - plausible*
Sue remembered the **letter** that the butler with the cup accidentally **shattered** ...

(d) *implausible - implausible*
Sue remembered the **letter** that the butler with the tie accidentally **shattered** ...

3.2 Method

As described in Section 2.1., in attention heads, each token allocates different amounts of attention to previous tokens depending on the relevance of the two tokens[2].

With such property of Transformers, the capacity of attention heads in detecting plausibility is measured with two terms that indicate how attention allocation patterns differ between plausible sentences and implausible sentences: *accuracy* and *attention difference*.

Accuracy indicates how likely the plausible noun is to get higher attention than the implausible noun in a certain attention head (See Equation (1)).

$$\text{Accuracy}_{\text{lh}} = \frac{\sum_{j=1}^{k}[\text{Attn}(\text{pl}_\text{j}, \text{v}_\text{j}) > \text{Attn}(\text{impl}_\text{j}, \text{v}_\text{j})]}{\text{k}} \quad (1)$$

where *lh* refers to the location of attention heads (*l* for the *l*th layer and *h* for the *h*th head in the *l*th layer), *j* refers to the sentence id, pl_j and $impl_j$ refer to the plausible and implausible nouns to be compared in the *j*th sentence set, v_j refers to the verb in the *j*th sentence, and *k* is the number of sentence sets.

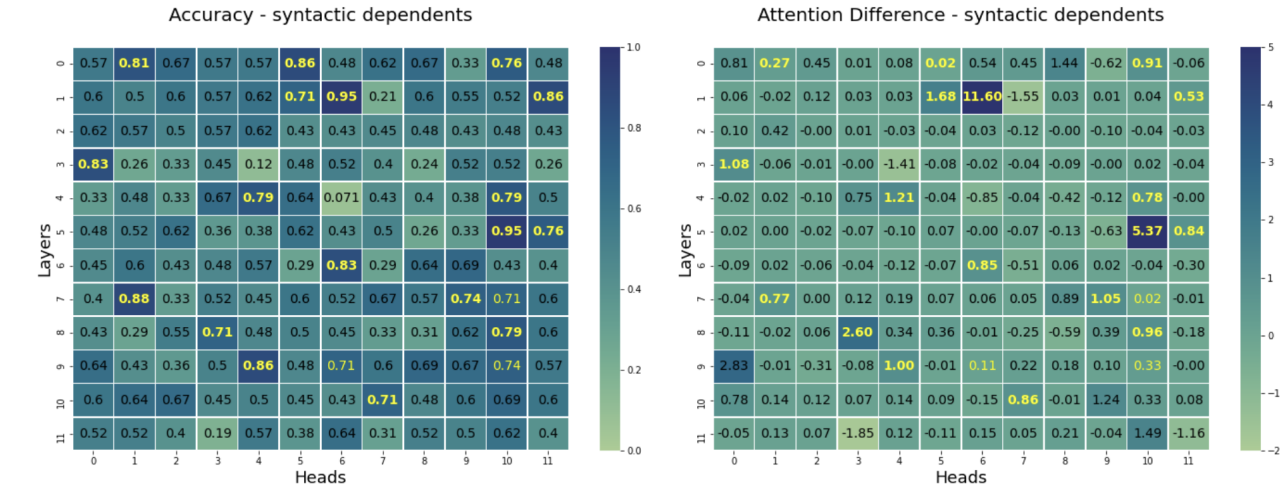
Attention Difference indicates how much more attention plausible nouns get compared to implausible nouns (See Equation (2)).

$$\text{AttnDiff}_{\text{lh}} = \sum_{j=1}^{k}[\text{Attn}(\text{pl}_\text{j}, \text{v}_\text{j}) - \text{Attn}(\text{impl}_\text{j}, \text{v}_\text{j})] \quad (2)$$

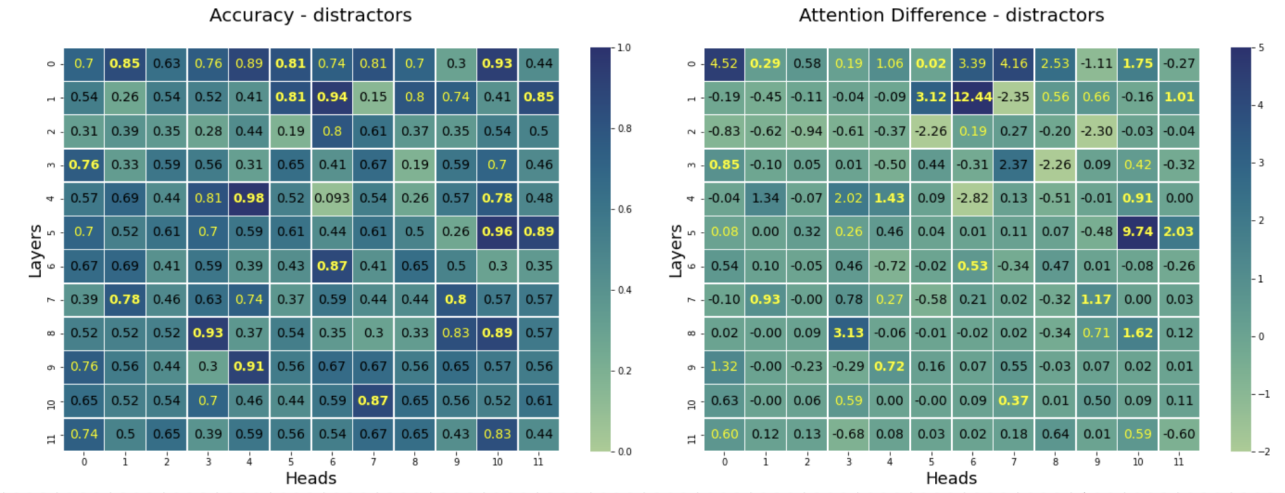
In order to ensure that the heads do not particularly work for tokens that form syntactic dependency but work for semantically related tokens, I will measure *accuracy* and *attention difference* not only with comparing attentions to plausible dependents and implausible dependents (*plate* vs. *letter* in (a)-(d)), but also with comparing attentions to plausible distractors and implausible distractors (*cup* vs. *tie* in (a)-(d)). By doing so, it is able to find attention heads that can judge the plausibility between nouns and verbs regardless of syntactic compatibility between them. Thus, for each set of sentences, there are four comparisons between

[1] In experiments with GPT2, 28 sets were used after removing sets of sentences whose tokens of interest cannot be recognized as a single token by the tokenizer.

[2] The relevance can be defined in terms of functions that attention heads serve. For instance, if an attention head is specialized for detecting *subject-verb* dependency relation, the amount of attention can reflect how likely two tokens are in the *subject-verb* relationship (Voita et al., 2019)



(a) Results from comparing plausible subjects and implausible subjects



(b) Results from comparing plausible subjects and implausible distractors

Figure 1: Accuracy and attention difference by attention heads. Attention heads annotated with bold-yellow showed accuracy greater than 0.70 in both subjects-comparison and distractors-comparison and thus considered to be specialized for plausibility processing; Attention heads annotated with non-bold-yellow are the ones that showed accuracy greater than 0.70 only for the corresponding condition; Attention heads annotated with black are found to be insensitive to plausibility (accuracies are less than 0.7 for both noun types).

plausible and *implausible* conditions: (pl-pl vs. pl-impl), (impl-pl vs. impl-impl), (pl-pl vs. impl-pl), (pl-impl vs. impl-impl), where the first corresponds to correct dependents and the second corresponds to distractors.

GPT2-small model (Radford et al., 2019) was used to extract attention values, which has 144 attention heads (12 heads per each of 12 layers). The pre-trained model and attention allocation patterns in each head will be accessed through HuggingFace (Wolf et al., 2019). I use this model for the rest of experiments in the present paper.

3.3 Results

Figure 1. shows the accuracy and the attention difference by attention heads. I consider attention heads are able to process plausible relationships between nouns and verbs if their accuracy in finding plausible nouns is greater than 70%. To select attention heads that can process the semantic plausibility regardless of the syntactic dependency relation between the noun and the verb, I consider attention heads whose accuracy is greater than 70% in both noun types (plausible vs. implausible syntactic dependents and plausible vs. implausible distractors).

With such criteria, eight-teen attention heads are recognized to be specialized for semantic plausibil-

ity: [(0, 1), (0, 5), (0, 10), (1, 5), (1, 6), (1, 11), (3, 0), (4, 4), (4, 10), (5, 10), (5, 11), (6, 6), (7, 1), (7, 9), (8, 3), (8, 10), (9, 4), (10, 7)], where the first numbers refer to indexes of layers and the second refer to indexes of heads (i.e., (*i*, *j*) refers to the *j*th head in the *i*th layer.))

Among the attention heads that are found to process semantic plausibility, two attention heads - (1, 6) and (5, 10) - especially show noteworthy performance in detecting plausible nouns given the high numbers of attention differences. Later in this paper, it will be discussed whether such high-performing attention heads necessarily make a greater contribution to Transformers' ability to process semantic plausibility.

3.4 Discussion

This section showed that a set of attention heads are specialized for semantic plausibility processing by showing their ability to determine which noun forms a semantically plausible relation with a certain verb. Such plausibility processing ability is found to be independent of their ability to process syntactic dependencies. To be specific, their ability to process plausibility is not limited to processing syntactic dependents of verbs, but it is also applicable to nouns that do not form any syntactic dependencies with verbs (i.e., distractors).

Unlike attention heads specialized for processing a certain syntactic relation and superficial linguistic information such as word position or word rarity is clustered in a relatively small region (Voita et al., 2019), it seems that the components that are specialized for semantic plausibility are relatively evenly distributed across twelve layers and take up are greater region: 18 attention heads out of 144 attention heads in GPT2-small model.

In the next section, it will be discussed how these plausibility processing attention heads exert causal effects on GPT2's plausibility processing ability.

4 Causal effects of plausibility-processing attention heads on GPT2's plausibility sensitivity

In the first experiment, attention heads capable of detecting plausible relations between nouns and verbs are found. However, causal effects from those attention heads to GPT2's general sensitivity to semantic plausibility still remain unanswered. In this section, how such attention heads influence Transformers' sensitivity to plausibility between nouns and verbs will be examined. To this end, I raise two questions: (1) How GPT2's responses to plausible/implausible verb-noun pairs changes when the models are with/without plausibility-processing attention heads? and (2) How does GPT2's plausibility-sensitivity change as attention heads are gradually pruned? Is the change continuous or gradual? The answers to these questions will be provided in the following experiments.

4.1 Influence of a set of plausibility-processing heads to plausibility sensitivity

In this section, I examine how GPT2's responses to plausible and implausible sentences change depending on whether the model is with a set of plausibility-processing heads or without them. To this end, I compare the models' responses with the human responses from Cunnings and Sturt (2018), considering that the model is more capable of processing plausibility if their responses are similar to the human response.

4.1.1 Method

In Cunnings and Sturt (2018), the degree of difficulty that people encounter while processing a certain plausible and implausible noun-verb pair was measured with reading times that are measured at verb (*shattered* in (a)-(d))[3]. To compare humans' responses with GPT2's, I compute surprisals (Hale, 2001; Levy, 2008), also measured at verbs, as a metric that represents processing difficulty of the language model, given a large set of evidence manifesting that surprisals computed from neural network language models can simulate human sentence processing patterns (Futrell et al., 2019; Michaelov and Bergen, 2020)[4]

Surprisal is a term that estimates the degree of the unexpectedness of tokens given their preceding context, which is computed by taking the negative log probability of a token conditioned on its preceding words (See Equation (3)). In neural network language models, the surprisal of a word is computed using the softmax-activated hidden state before consuming the word (Wilcox et al., 2018).

$$\text{Surprisal}(w) = -\log_2 P(w|h) \qquad (3)$$

[3]The original paper also talks about the spillover region after the verbs of interest, but this study focuses on the reading times measured at the verb.

[4]A large set of previous studies showed that surprisals computed with neural language models are highly correlated with human reading times

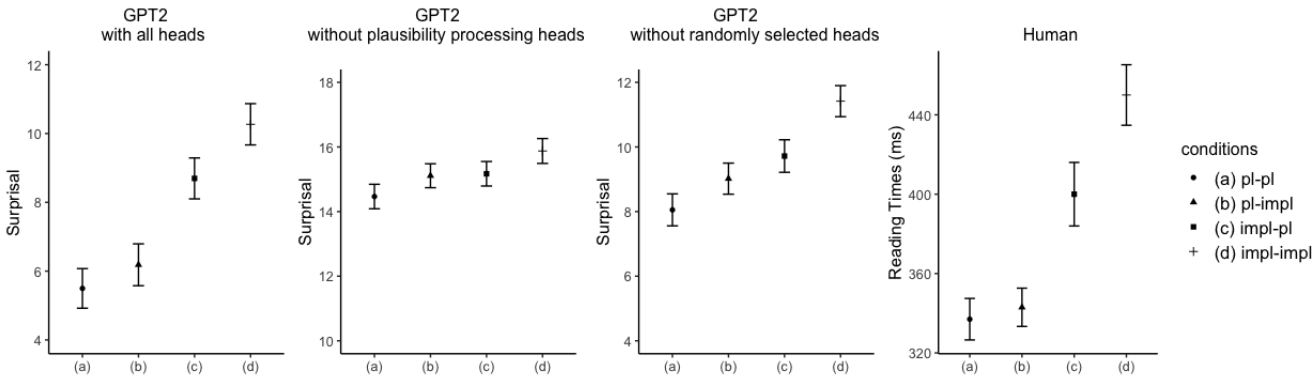


Figure 2: Surprisals computed from GPT2s with different sets of attention heads and reaction times from human subjects for processing different types of noun-verb pairs. Human reading times are obtained from Cunnings and Sturt (2018). Shapes at the center and intervals for each condition represent means and standard errors.

where h is the softmax-activated hidden state of the sentence before encountering the current word.

Both reading times and surprisals measured at verbs of interest are expected to be greater in sentences with implausible nouns than in sentences with plausible nouns since it is less likely for humans and language models to anticipate a certain verb after encountering a noun that is in an implausible relationship with the verb.

The sets of attention heads in GPT2 that are examined are as follows: i) all 144 attention heads in GPT2-small model, ii) GPT2 without plausibility-processing heads, iii) GPT2 with the same number of attention heads as ii), but the heads to prune selected randomly. I included the third model in order to see whether the effect of removing plausibility processing attention heads is simply caused by taking away information in GPT2 or it is caused by specifically removing plausibility processors. In order for reliability, we used 100 different random attention head sets for iii), and computed the average of surprisals from 100 models for each sentence. Attention heads were pruned by replacing attention values with zeros, following Michel et al. (2019).

4.1.2 Results

Surprisals computed from GPT2 models that are with different sets of attention heads and reaction times from human subjects when processing different types of noun-verb pairs are shown in Figure 2.

With the GPT2 model that is with the entire set of attention heads, it is shown that the way the model process plausibility of noun-verb pairs are similar to the way humans do: i) significantly lower processing difficulties are found when syntactic dependents are in a plausible relation with the verb than when they are in an implausible relation and ii) plausibility effects are found even with nouns that do not form syntactic dependency with the verb, (i.e., processing difficulties are greater in (b) and (d) than in (a) and (c)), though the plausibility effects are much smaller than the cases where nouns are in the syntactic relation with verbs. Plausibility effects observed for distractors in GPT2 and humans are due to the illusion of plausibility (Cunnings and Sturt, 2018): even distractors that cannot build syntactic dependency with cues (verbs) can be illusorily considered as the syntactic dependents, causing moderate plausibility effects while sentence processing.

Then, how do the eight-teen attention heads that are found to be specialized for plausibility processing in the previous section contribute GPT2's ability to simulate human responses in plausibility processing? The second graph in Figure 2. shows that the differences in surprisal by conditions become much smaller when the model is without those attention heads.

Importantly, such a decrease is not likely to be the effect that is caused by simply removing components in GPT2. In the third graph, it is shown that when randomly selected eight-teen attention heads are removed the GPT2 model better simulates human responses in processing plausibility than the model whose pruned attention heads are specifically specialized for plausibility processing. This supports that the plausibility-processing attention heads are making an exclusive contribution to GPT2's ability to process plausibility.

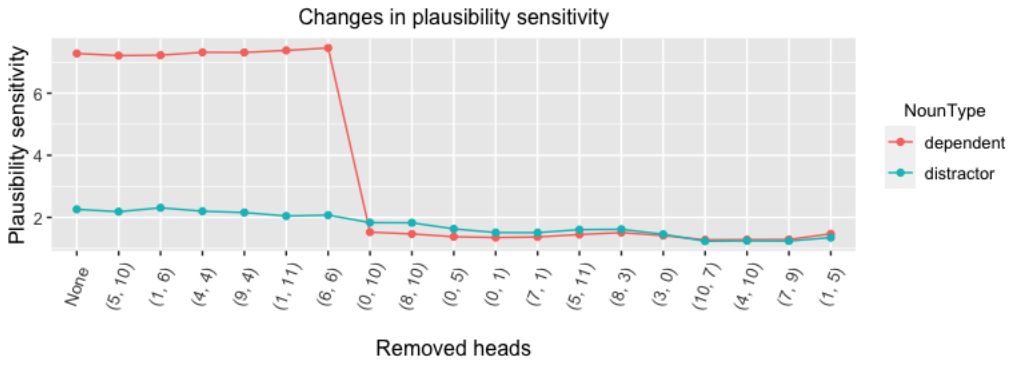


Figure 3: Changes in plausibility sensitivity by noun types as attention heads are gradually pruned. Attention heads are plausibility-processing attention heads, and they are ordered by accuracies in determining plausible nouns over implausible nouns.

4.2 Gradual changes in GPT2's plausibility sensitivity as attention heads are pruned

The previous section examined the influence of the set of attention heads specialized for plausibility processing on GPT2's ability to process plausibility. Though it was shown that plausibility-processing attention heads collectively contribute to GPT2's ability to process plausibility unlike other sets of attention heads, it is unanswered how individual attention heads contribute to GPT2's plausibility processing ability. Do they have fairly balanced contributions to GPT2's ability to process plausibility? Or, only a small set of plausibility-processing attention heads are enough to account for most of the plausibility-processing ability of GPT2?

In order to answer these questions, the following experiment investigates how GPT2's general sensitivity to plausibility gradually changes as attention heads are pruned one by one.

4.2.1 Method

This study operationalizes GPT2's plausibility sensitivity as the difference in *surprisals* measured at the verbs of interest ('*shattered*' in (a)-(d)) in sentences with plausible nouns and in sentences with implausible nouns as shown in Equation (4).

$$\begin{aligned}&\text{PlausibilitySensitivity} = \\ &\quad \text{surprisal}_{\text{impl}}(verb) - \text{surprisal}_{\text{pl}}(verb)\end{aligned} \tag{4}$$

, where $\text{surprisal}_{\text{pl}}$(verb) and $\text{surprisal}_{\text{impl}}$(verb) refer to surprisals measured at the verb in a sentence with a plausible noun and in a sentence with an implausible noun, respectively.

I computed two plausibility sensitivities: one that compares surprisals at verbs when having plausible syntactic dependents of verbs in sentences and having implausible syntactic dependents ({(c)+(d)} - {(a)+(b)}) and the other that compares surprisals when having plausible distractors of verbs and implausible distractors ({(b)+(d)} - {(a)+(c)}).

Both types of plausibility sensitivities are measured at each point after gradually removing a plausibility processing attention head one by one.

Attention heads were pruned in order of their accuracies[5] in detecting plausible nouns over implausible nouns.

4.2.2 Results

Figure 3. plots how the plausibility sensitivities for both types of noun-verb relations change as plausibility-processing attention heads are removed gradually.

When it comes to the relation between distractors and verbs, the changes in plausibility sensitivity seem to be continuous. Such patterns suggest that the set of plausibility processing attention heads make a collective contribution to plausibility effects in the distractor and verb relation.

In contrast, plausibility sensitivity for the relation between syntactic dependent and verb shows a drastic decrease upon the removal of the attention head (0, 10). The effect from the removal of the head (0, 10) shows that this particular head exerts a huge amount of causal effects on GPT2's general sensitivity to plausible relations between syntactic subjects and verbs. Figure 4 confirms that the head (0, 10) can cause a huge amount of causal contribution on GPT2's plausibility processing ability since it accounts for a great portion of plausibility effects in Transformers, if not all.

[5] I used the average values of the accuracies for syntactic dependents and for distractors that were computed in Section 3.

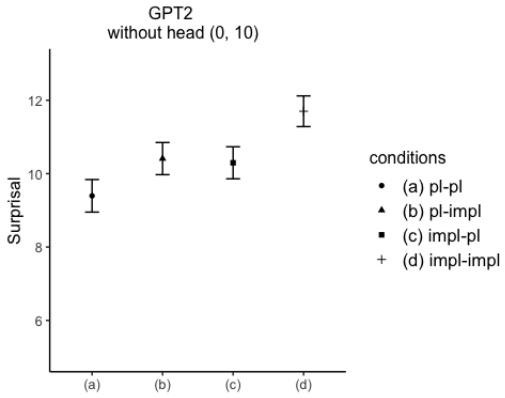


Figure 4: Surprisals by conditions computed with the GPT2 without a single attention head (0, 10). Shapes at the center and intervals for each condition represent means and standard errors, respectively.

4.3 Discussion

Results of this section suggest plausibility processing in GPT2 requires a collective contribution from a large set of plausibility processing attention heads, given that plausibility sensitivity decreases continuously as attention heads are gradually pruned. At the same time, however, it was shown that the amount of causal effects that each attention head makes are highly imbalanced because the attention head (0, 10) leads to a much more drastic decrease in plausibility sensitivity than other heads. Taken together, although a single attention head can account for a great portion of the plausibility effects, other plausibility-processing attention heads make an additional contribution to GPT2's plausibility-processing ability.

Interestingly, the head (0, 10) did not achieve the best performance in detecting plausible nouns over implausible nouns in the previous experiment. Further investigation needs to be conducted to show what properties of this particular attention head would lead such a huge amount of causal effects on GPT2's plausibility sensitivity. More importantly, this suggests that analyzing the causal effects each attention head makes is indispensable to understanding the role that attention heads serve, provided that performance that each attention head shows in processing particular linguistic information does not necessarily lead to the eventual contribution to the model's performance.

The fact that the changing patterns in plausibility sensitivity are different by the type of syntactic role that nouns serve (syntactic dependent or distractor) also urges further research on how the plausibility processing attention heads affects Transformers' general processing ability needs to be understood in relationship to other attention heads, especially the ones specialized for syntactic relation.

One additional interesting finding is that the level of surprisals from GPT2 without plausibility processing heads is much higher than other models. For instance, surprisals in condition (a) increase by around nine bits after removing the plausibility processing attention heads (Compare the first two graphs in Figure 2). This indicates removing plausibility processing attention heads causes serious harm to GPT2's general ability to predict the next token given the preceding context. I suppose the plausible-processing attention heads have developed to have other key functions that relate to sentence comprehension in addition to the ability to process plausibility. Since this topic is beyond the scope of this paper, I would leave this question to future studies.

5 Conclusion

In this paper, a set of experiments showed that a number of attention heads, which are diffusely distributed across layers in Transformers, can process plausible relations between nouns and verbs. Moreover, it was observed that they make imbalanced but collective causal contributions to Transformers' human-like way ability to process plausibility, which establishes the importance of causal effect analysis in attention-head-probing-studies.

Although the results provide a window into how Transformers process semantic knowledge of plausibility, this study has a few limitations to be addressed. First, the scope of the study is restricted to the plausible relation between nouns and verbs although there exist many different types of semantic knowledge. In order to test the generalizability, the scope of the study needs to be extended. Also, it does not explain how the plausibility processing attention heads interact with other components in the model. However, such information would be a key to a deeper understanding of the roles of plausibility processing attention heads since it could explain the mechanism of how the attention heads contribute to Transformers' human-like ability to process plausibility.

With these limitations addressed, I anticipate further improvements in explaining Transformer's plausibility processing ability will be made on the basis of findings in the present study.

References

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

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Sudeep Bhatia and Russell Richie. 2022. Transformer networks of human conceptual knowledge. *Psychological Review*.

Sudeep Bhatia, Russell Richie, and Wanling Zou. 2019. Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, 29:31–36.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Ian Cunnings and Patrick Sturt. 2018. Retrieval interference and semantic interpretation. *Journal of Memory and Language*, 102:16–27.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

Simon Jerome Han, Keith Ransom, Andrew Perfors, and Charles Kemp. 2022. Human-like property induction is a challenge for large language models.

Jae-young Jo and Sung-Hyon Myaeng. 2020. Roles and utilization of attention heads in transformer-based neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3404–3417.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Danny Merkx and Stefan L Frank. 2020. Human sentence processing: Recurrence or attention? *arXiv preprint arXiv:2005.09471*.

James A Michaelov and Benjamin K Bergen. 2020. How well does surprisal explain n400 amplitude under different experimental conditions? *arXiv preprint arXiv:2010.04844*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring bert's sensitivity to lexical cues using tests from semantic priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2021. Do language models learn typicality judgments from text? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. 2022. Copen: Probing conceptual knowledge in pre-trained language models. *arXiv e-prints*, pages arXiv–2211.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Sello Ralethe and Jan Buys. 2022. Generic overgeneralization in pre-trained language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3187–3196.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*.

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*.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do rnn language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.