As TACL action editor for submission 1815, "Theoretical Limitations of Self-Attention in Neural Sequence Models",   I am happy to tell you that I am accepting your paper subject (conditional) to your making specific revisions within two months.

While all three reviewers found the work timely, worthwhile and well executed, the more CL-oriented reviewers also raised issues with the clarity of some of the arguments, and with (the lack of) explicit discussion connecting the theoretical results to practical language understanding work and in particular the empirical success of the self attention architecture. As the main audience of this journal are the CL/NLP community, the paper should cater more to this crowd and make some things more explicit. In particular, see the list of mandatory revisions below. You are also encouraged to carefully read the reviews and identify additional revisions that could strengthen the paper.

LIST OF MANDATORY REVISIONS:

(a) Make more explicit the connection between your presentation to the one in the original Vaswani et al transformer paper (queries, keys and values), or otherwise refer to other common and somewhat standardized presentations of the transformer model that do not use this terminology.

(b) Add more salient discussion of skip-connection / residual connections, and in particular remark on them in the proofs. While reviewer C believes the results should hold also with residual connections, I initially shared reviewer A's reaction of not understanding this detail, and I am fairly certain other readers will also face similar confusion. Please enhance the relevant proofs with discussion of why they work also in the presence of residual connections.

(c) Expand the discussion on the relevancy of the result. In particular:
- (c1) Expand on your thoughts on what the value of showing such an asymptotic property is (traditionally, proofs of unexpressability in the limit are really stand-ins for proofs of ungeneralizability).
- (c2) In particular, expand on the applicability of the results to processing of natural language data and what is the takeaway to NLP practitioners, if any. Empirical evidence suggests that the transformer architecture is indeed preferable to other architectures in capturing longer distance context, how is this reconciled with the theoretical findings?
- Note that it is perfectly OK to have a theoretical result that does not have any relevance to empirical NLP work, and I find the result to be interesting enough to be published also without such relevance. But, if that is indeed the case, it needs to be stated explicitly.

(d) Attempt to Improve / simplify / expand the presentation of mathematics used in sections 5 and 6 to cater to audiences who are less familiar with these literature and techniques.

Generally, your revised version will be handled by the same action editor (me) and the same reviewers (if necessary) in making the final decision --- which, *if* all requested revisions are made, will be final acceptance.

You are allowed one to two extra pages of content to accommodate these revisions. To submit your revised version, follow the instructions in the "Revision and Resubmission Policy for TACL Submissions" section of the Author Guidelines at https://transacl.org/ojs/index.php/tacl/about/submissions#authorGuidelines .

Thank you for submitting to TACL, and I look forward to your revised version!

Yoav Goldberg
Bar Ilan University/Allen Institute for Artificial Intelligence
yoav.goldberg@gmail.com

--
--
....THE REVIEWS....
--
--

Reviewer A:

APPROPRIATENESS: Does the paper fit in TACL? (Please answer this question in light of the desire to broaden the scope of areas represented in the ACL community.):

5. Certainly.

CLARITY: For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?:

4. Understandable by most readers.

INNOVATIVENESS: Does this paper break new ground in topic or content? How exciting and innovative is the research?

Note that a paper can score high for innovativeness even if its impact will be limited.
:

4. Creative: An intriguing theoretical result or analysis that is substantially different from previous research.

SOUNDNESS/CORRECTNESS: First, is the theoretical approach sound and well-chosen? Second, can one trust the claims of the paper -- for example, are they supported by an appropriate proof or analysis?:

1. Fatally flawed.

RELATED WORK: Does the submission make clear where the work sits with respect to existing literature? Are the references adequate? Are the benefits of the theoretical approach well-supported?

Note that the existing literature includes preprints, but in the case of preprints:

- Authors should be informed of but not penalized for missing very recent and/or not widely known work.
- If a refereed version exists, authors should cite it in addition to or instead of the preprint.

:

5. Precise and complete comparison with related work.

SUBSTANCE: Does this paper have enough substance (in terms of the amount of work), or would it benefit from more ideas or analysis?

Note that papers or preprints appearing less than three months before a paper is submitted to TACL are considered contemporaneous with the submission. This relieves authors from the obligation to make detailed comparisons that require additional experiments and/or in-depth analysis, although authors should still cite and discuss contemporaneous work to the degree feasible.

:

4. Represents an appropriate amount of work for a publication in this journal. (most submissions)

IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the ideas are novel, will they also be useful or inspirational? If the results are sound, are they also important? Does the paper bring new insights into the nature of the problem?:

4. Some of the ideas or results will substantially help other people's ongoing research, or the paper is otherwise important.

IMPACT OF PROMISED SOFTWARE: If the authors state (in anonymous fashion) that their software will be available, what is the expected impact of the software package?:

1. No usable software released (a fine score for a submission where the main contribution is theoretical).

IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion) that datasets will be released, how valuable will they be to others?:

1. No usable datasets submitted (a fine score for a submission whose main contribution is theoretical).

TACL-WORTHY AS IS? In answering, think over all your scores above. If a paper has some weaknesses, but you really got a lot out of it, feel free to recommend it. If a paper is solid but you could live without it, let us know that you're ambivalent.

Reviewers: after you save this review form, you'll have to make a confidential recommendation to the editors via pull-down menu as to: what degree of revision would be needed to make the submission eventually TACL-worthy?

:

1. Poor: I'd fight to have it rejected.

Detailed Comments for the Authors

Reviewers, please draft your comments on your own filesystem and then copy the results into the text-entry box. You will thus have a saved copy in case of system glitches.
:

This paper takes a formal approach to the analysis of languages generatable by Transformer encoders, a particular configuration of feed-forward neural networks equipped with attentional connections. (It should be noted that 'transformer' as originally presented refers to the encoder-decoder architecture but it seems plain from this paper that only the encoder is considered, which is not inappropriate; there is plenty of justified interest in the language modeling properties of the encoder,) This work mirrors an earlier era's
fundamental analysis of formalisms in the chomsky hierarchy and beyond, which stretched into the earlier days of neural networks when similar analysis was conducted for these machines. The results of the paper are quite surprising: Transformers, despite appearing to be far better in practice at capturing various aspects of human language in more efficient and more complete ways than other neural network models (RNN, CNN) and other formalisms (TAG, CFG, FSA), are in fact limited in their ability to model some finite-state and context-free languages, which have been shown to be less expressive than what is needed to capture natural language.

The paper is very well written and the design and argumentation behind the central proofs is compelling and even captivating (for the right audience). The references appear to be quite comprehensive and seat the work appropriately. I am in general in favor of this kind of work and all things being equal would like to champion it above SOTA-chasing incremental advances that may or may not be replicable or applicable outside an exceedingly narrow case. However, there appears to be a fatal flaw in the argumentation of this paper. There also appears to be a minor flaw in part of the informal argument. I remain wary of my own misunderstanding of the key conclusions and arguments, and admit that fully understanding the lowest level of proof detail required more time than I was able to commit, however it is my belief that the fatal flaw obviated the need to read further. I welcome a response from the authors or co-reviewers since I am only somewhat confident of my finding.

The central argument of the paper is that if a certain portion of inputs to a transformer is restricted in a particular way, the transformer becomes sensitive in its prediction to a subset of the remaining inputs. As some finite-state and context free languages must be sensitive to every value of every input, the transformer thus cannot adequately discriminate between in-language and out-language inputs.

(Minor issue): I am confused by the discussion of languages that can and can't be shown to be modeled by transformers in lines 420-487. It is said "A crucial difference between these languages and PARITY / 2DYCK is that fixing a few inputs can easily force nonmembership, e.g. a single 0 for 1*, and an a in the second half for a^n b^n." Couldn't you use the same argument to restrict the first element to ']' and thereby force nonmembership in 1DYCK (and thus 2DYCK)? I do buy this argument for PARITY; clearly you can set C such that even when forcing Cn input symbols to any particular value, the status of the string is unknown. If I understand the general argument as applied to PARITY, then you could force 1-Cn bits to be all 0. If done, then for all sequences longer than 2*(1-Cn+c) PARITY is unrecognizable; you could at least test non-membership for shorter sequences (though still not membership).

The key theorem is that a restriction, if chosen properly, ultimately limits the number of inputs to a constant. As the input length grows, more of the input is ignored, and this proves fatal to the recognition of certain languages (see above).

(Major issue):If I properly understand the idea behind the proof of the depth reduction lemma that makes up the core of the proof of the key theorem (it is possible that I don't) then the mechanism by which a bounded number of sensitive inputs is obtained is by finding those input positions and values that yield maximum attention, fixing them as needed, and thereby overwhelming other inputs, rendering them insensitive. I did not follow the detailed justification for why this is possible and why this is bounded too closely, because I believe a fundamental flaw in the understanding of the transformer model ultimately dooms this strategy, as follows:

The Transformer paper itself (section 3.1) critically notes the inclusion of a so-called residual connection at each layer that ensures an aspect of each input is propagated up the layer hierarchy, independent of the attention mechanism. This means that each unit will be dependent on the corresponding unit (and ultimately input symbol) of the layer below it. The restriction selection mechanism does not take this into account. While it seems appropriate to abstract away specifics of attention design, representation, and even to treat soft attention as hard, the residual connection is a critical part of the model and cannot be ignored. That there is a required influencer for each node and that these cover ultimately the entire input sequence would seem to undermine the core argument of this paper.

Setting aside the possible flaw noted above, it is important to ask, is there a practical effect of the results in this paper? The restrictions that accomplish the described insensitivity behavior are not arbitrary, but are in fact quite carefully crafted. This makes the result in some sense akin to the adversarial image recognition findings, which showed brittle recognition properties such that an imperceptible (to human) change in an image could lead to catastrophic recognition results. But that brittleness is based on training methodology, not formalism, and transformers, as well as RNNs, CNNs, and even pre-neural models (and pre-statistical models) are sucseptible to that or similar brittleness. The formalism brittleness would be much harder to trigger in practice (since access to the typically unexposed weights is necessary, in order to find maximal attention units), so the threat
of adversaries is not increased. The degree to which Transformer's (formal) limits clash with human language expressivity does not seem to be, in practice, more severe than with other formalisms; indeed, it seems empirically to be less severe. The fundamental problem of full human language modeling remains in any case; it requires wide ranging, sporadic, multimodal, though ultimately finite context to capture all of common sense. Transformer *does* appear to capture more, longer-distance context and can in practice be trained to do so better than other models.

REVIEWER CONFIDENCE:
3. Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math, experimental design, or novelty.

Reviewer B:

APPROPRIATENESS: Does the paper fit in TACL? (Please answer this question in light of the desire to broaden the scope of areas represented in the ACL community.):
3. Unsure.

CLARITY: For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?:
2. Important questions were hard to resolve even with effort.

INNOVATIVENESS: Does this paper break new ground in topic or content? How exciting and innovative is the research?

Note that a paper can score high for innovativeness even if its impact will be limited.
:
3. Respectable: A nice research contribution that represents a notable extension of prior work.

Note that a paper could score high for originality even if the results do not show a convincing benefit.

SOUNDNESS/CORRECTNESS: First, is the theoretical approach sound and well-chosen? Second, can one trust the claims of the paper -- for example, are they supported by an appropriate proof or analysis?:

4. Generally solid work, although there are some aspects of the approach I am not sure about or the analysis could be stronger.

RELATED WORK: Does the submission make clear where the work sits with respect to existing literature? Are the references adequate? Are the benefits of the theoretical approach well-supported?

Note that the existing literature includes preprints, but in the case of preprints:

- Authors should be informed of but not penalized for missing very recent and/or not widely known work.
- If a refereed version exists, authors should cite it in addition to or instead of the preprint.

:

4. Mostly solid bibliography and comparison, but there are a few additional references that should be included.

SUBSTANCE: Does this paper have enough substance (in terms of the amount of work), or would it benefit from more ideas or analysis?

Note that papers or preprints appearing less than three months before a paper is submitted to TACL are considered contemporaneous with the submission. This relieves authors from the obligation to make detailed comparisons that require additional experiments and/or in-depth analysis, although authors should still cite and discuss contemporaneous work to the degree feasible.

:

4. Represents an appropriate amount of work for a publication in this journal. (most submissions)

IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the ideas are novel, will they also be useful or inspirational? If the results are sound, are they also important? Does the paper bring new insights into the nature of the problem?:

3. Interesting but not too influential. The work will be cited, but mainly for comparison or as a source of minor contributions.

IMPACT OF PROMISED SOFTWARE: If the authors state (in anonymous fashion) that their software will be available, what is the expected impact of the software package?:

1. No usable software released (a fine score for a submission where the main contribution is theoretical).

IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion) that datasets will be released, how valuable will they be to others?:

1. No usable datasets submitted (a fine score for a submission whose main contribution is theoretical).

TACL-WORTHY AS IS? In answering, think over all your scores above. If a paper has some weaknesses, but you really got a lot out of it, feel free to recommend it. If a paper is solid but you could live without it, let us know that you're ambivalent.

Reviewers: after you save this review form, you'll have to make a confidential recommendation to the editors via pull-down menu as to: what degree of revision would be needed to make the submission eventually TACL-worthy?
:

3. Ambivalent: OK but does not seem up to the standards of TACL.

Detailed Comments for the Authors

Reviewers, please draft your comments on your own filesystem and then copy the results into the text-entry box. You will thus have a saved copy in case of system glitches.
:

Summary
This paper present a theoretical study of the expressive power of (an idealization of) the Transformer model. It presents two formal languages, 2Dyck and Parity, as key diagnostic learning targets, and finds they are beyond reach for a version of the Transformer based on ‘hard attention’, as well as for a version based on ‘soft attention’.

Main contributions are (1) Theorem1/Corrolary2/Lemma4, and their proofs (using and adapting an approach from Boolean circuit literature from the 1980s), that together show that a Transformer with hard attention cannot model Dyck2 & Parity; (2) Lemma5/Theorem6, and their proofs (using Lipschitz continuity), that together show that a Transformer with soft attention cannot learn non-trivial probability distributions over Dyck2 & Parity.

Assessment
The paper is generally well written, with a good general introduction and accessible discussion, and the nice habit to provide intuitions behind the proofs before diving into the math. Moreover, it is very good to see a ‘theory-paper’ in a field dominated by computational experiments without much theoretical motivation. Nevertheless, although I don’t dispute the validity of the main theorems, I am not convinced about their relevance, and furthermore, must admit that I have been unable to completely follow the proofs so I cannot attest to the correctness of the presented work.

I’ll try to explain why the main claims make intuitive sense to me, before detailing where my confusion starts and why I have doubts about their relevance. 2Dyck & Parity are formal languages, constructed so that the validity of a symbol at each position in a string depends on symbols at arbitrary distance away in the string. In the Transformer, attention heads track dependencies between 2 symbols at a time, and aggregate such information step by step in each next layer. With a finite number of attention heads and a finite number of layers, it makes sense to me that one can design formal languages that break these finite limits (cf Dehghani et al).

Confusion
A formal proof to replace such intuitions would be progress. But my confusion about the formalization already starts on page 3, where I can’t square the definitions of (hard and soft) attention heads with those in Vaswani et al 2017. Where are the value, key and query components that are so crucial in that work?

The confusion continues throughout section 5, where first ‘restrictions’ are introduced. The basic idea, the paper says, is that a small fraction of the input is fixed in a particular way. Perhaps this is a round-about way to define a subset of the Parity/2Dyck languages for which the Tranformer’s failure will be demonstrated. Restrictions are then said to be ‘applied to the transformer’. That is also counterintuitive; I would expect restrictions on models in proofs of what model can do (even when restricted), and not in proofs on what they cannot do. It is probably my lack of knowledge about the Boolean circuit literature, but I would need more explanation and perhaps terminology more carefully tuned to an ACL*-audience here to follow this entire section.

I don’t fare much better in section 6, where a different set of mathematical techniques is used to prove theorem 6 and lemma 5. The paper uses a simple PCFG to define the input language, but I am immediately at a loss when the proofs start after giving theorem 6. What is an *equally likely* non-member that yields similar but different output activations?

All this might be due to my limited knowledge of the mathematics that is applied here. I leave it to the editor and other reviewers to assess whether the general reader of TACL will do better, or whether the paper really needs to be rewritten to make it more selfcontained.

Relevance
Even if all the proofs are correct, I have some doubts about how relevant all this is for understanding the behavior of the Transformer on real language data. The paper, on page 10, briefly discusses the very reasonable objection that the Transformer circumvents limitations of the sort studied in this paper by using a large number of layers and attention heads, and that natural language data simply doesn’t contain those deeply nested dependencies, because humans fail at them too. This is such a crucial issue for assessing the relevance of the presented proofs, that I think it would deserve some longer discussion and some supportive empirical data.

Recommendation
If there are other reviewers that have been able to check the proofs, this paper might be publishable after some editing that clarifies the relation with the definition of attention in Vaswani et al., improves the explanation

of the mathematics used in sections 5 and 6, and more strongly argues for the empirical relevance of the results.

REVIEWER CONFIDENCE:
2. Willing to defend my evaluation, but it is fairly likely that I missed some details, didn't understand some central points, or can't be sure about the novelty of the work.

--

--
Reviewer C:

APPROPRIATENESS: Does the paper fit in TACL? (Please answer this question in light of the desire to broaden the scope of areas represented in the ACL community.):
5. Certainly.

CLARITY: For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?:
5. Very clear.

INNOVATIVENESS: Does this paper break new ground in topic or content? How exciting and innovative is the research?

Note that a paper can score high for innovativeness even if its impact will be limited.
:
3. Respectable: A nice research contribution that represents a notable extension of prior work.
Note that a paper could score high for originality even if the results do not show a convincing benefit.

SOUNDNESS/CORRECTNESS: First, is the theoretical approach sound and well-chosen? Second, can one trust the claims of the paper -- for example, are they supported by an appropriate proof or analysis?:
5. The theoretical approach is very apt, and the claims are convincingly supported.

RELATED WORK: Does the submission make clear where the work sits with respect to existing literature? Are the references adequate? Are the benefits of the theoretical approach well-supported?

Note that the existing literature includes preprints, but in the case of preprints:
- Authors should be informed of but not penalized for missing very recent and/or not widely known work.
- If a refereed version exists, authors should cite it in addition to or instead of the preprint.

:
5. Precise and complete comparison with related work.

SUBSTANCE: Does this paper have enough substance (in terms of the amount of work), or would it benefit from more ideas or analysis?

Note that papers or preprints appearing less than three months before a paper is submitted to TACL are considered contemporaneous with the submission. This relieves authors from the obligation to make detailed comparisons that require additional experiments and/or in-depth analysis, although authors should still cite and discuss contemporaneous work to the degree feasible.
:
4. Represents an appropriate amount of work for a publication in this journal. (most submissions)

IMPACT OF IDEAS OR RESULTS: How significant is the work described? If the ideas are novel, will they also be useful or inspirational? If the results are sound, are they also important? Does the paper bring new insights into the nature of the problem?:
3. Interesting but not too influential. The work will be cited, but mainly for comparison or as a source of minor contributions.

IMPACT OF PROMISED SOFTWARE: If the authors state (in anonymous fashion) that their software will be available, what is the expected impact of the software package?:
1. No usable software released (a fine score for a submission where the main contribution is theoretical).

IMPACT OF PROMISED DATASET(S): If the authors state (in anonymous fashion) that datasets will be released, how valuable will they be to others?:
1. No usable datasets submitted (a fine score for a submission whose main contribution is theoretical).

TACL-WORTHY AS IS? In answering, think over all your scores above. If a paper has some weaknesses, but you really got a lot out of it, feel free to recommend it. If a paper is solid but you could live without it, let us know that you're ambivalent.

Reviewers: after you save this review form, you'll have to make a confidential recommendation to the editors via pull-down menu as to: what degree of revision would be needed to make the submission eventually TACL-worthy?
:
4. Worthy: A good paper that is worthy of being published in TACL.

Detailed Comments for the Authors

Reviewers, please draft your comments on your own filesystem and then copy

the results into the text-entry box. You will thus have a saved copy in
case of system glitches.
:

The paper studies the ability of transformers to model formal languages, in particular PARITY and 2DYCK (representatives of regular and context-free languages, respectively). It is proven that with hard attention transformers, for sufficiently long input sequences, a (relatively) small number of input symbols can be fixed such that the output ignores other symbols, and thus cannot model PARITY or 2DYCK. For soft attention, a somewhat weaker limitation is established --- transformers converge to chance level when predicting the next symbol of a very long PARITY or 2DYCK sequence. The results demonstrate inferiority of transformers compared to RNNs and LSTMs in modelling formal languages (the latter can capture PARITY and 2DYCK). Given the empirical success of transformers, this puts in question the relevance of typical formal models from theoretical linguistics to natural language processing.

Before providing my feedback on this work, I would like to stress that my area of expertise is theoretical machine learning, not computational linguistics. Accordingly, my evaluation in terms of significance and novelty is based primarily on the text written by the authors, and should be taken with a grain of salt.

That said, I found the paper interesting, very well written, and technically solid (I did not verify every single detail in the proofs of Lemmas 4 and 5 but nonetheless believe the analysis is firm). It addresses a timely topic, and despite my critique (see below), I believe it does shed some light on the inherent differences between transformers and popular recurrent models (RNN/LSTM).

The most significant drawback I see in this work is that its results are asymptotic, i.e. apply only to input sequences that are very long. This means that from a practical perspective, it is unclear if the limitations established are at all relevant. The authors acknowledge this fact and discuss it explicitly in Section 7. Nonetheless, I would suggest being a bit more transparent in terms of the input lengths needed for the results to hold. As far as I could tell, there is an exponential and even a tetration ("exponential tower") blow-up. The second main comment I have is that it seems to me like the results for soft attention can actually be applied to any stateless model processing arbitrarily long inputs with a fixed number of parameters. This means that they do not shed light on the specific nature of transformers (self attention). Again, the authors mention this but I believe it could be stressed a bit more boldly (e.g. in the discussion in Section 7). Overall, the two shortcomings I raised are inherent to this work, i.e. cannot be addressed without major changes, and since they are acknowledged in the text, I do not think they should be treated as an impediment for publication.

Minor comments:
* Typo in line 151: agreeement

* Typo in line 195: reccurent
* Typo in line 355: y^k_(j) (parenthesis should be in superscript)
* In Theorem 1, it should be stated that the result applies to hard attention.
* In Equation (4), there is a typo with n being a subscript where it should be a superscript
* Typo in line 850: some some

REVIEWER CONFIDENCE:

2. Willing to defend my evaluation, but it is fairly likely that I missed some details, didn't understand some central points, or can't be sure about the novelty of the work.
