Supplement: Theoretical Limitations of Self-Attention in Neural Sequence Models

Michael Hahn

May 5, 2021

Here I am providing two supplements to the published TACL paper: First, a more formal writeup of the hard attention proof. This has benefited a lot from discussions with Gail Weiss and Will Merrill. Second, I am providing a missing detail in the soft attention proof (thanks for Navin Goyal and Satwik Bhattamishra for spotting this).

S1 Results for Hard Attention

**Theorem 1.** Let any hard attention transformer be given, and let $C \in (0, 1)$. Then there is a restriction $\rho$ and an integer $c > 0$ such that

$$|\{i \leq n : \rho_n(i) = \ast\}| \geq Cn$$

(for all sufficiently large $n$) and such that the function computed by the transformer on the restricted input depends only on $\leq c$ inputs, independent of input length $n$.

**Definition 2 (c-Transformer).** Let $c$ be a positive integer. A $c$-transformer is one in which the layer-0 activations $y_j^{(0)}$ depend on the embeddings not just at one position $j$, but are a function of the embeddings at $\leq c$ input positions:

$$y_j^{(0)} = f_{n,j}^{\text{inp}}((v_{i_1}^{(n)}, p_{i_1}^{(n)}), \ldots, (v_{i_c}^{(n)}, p_{i_c}^{(n)}))$$

for some indices $i_s^{(n)} \in \{1, \ldots, n\} (s = 1, \ldots, c)$.

**Definition 3.** We say $\rho' \succ \rho$ if, whenever $\rho'_n(i) = \ast$, then $\rho_n(i) = \ast$.

We write $\rho T$ for the function resulting from applying $\rho$ to $T$.

We write $\rho \Sigma^*$ for the set of inputs compatible with $\rho$.

With this technical notion, we show that we can reduce layers, iteratively removing the lowest layer until no self-attention layer is left:

**Lemma 4 (Depth Reduction Lemma).** Given a $c$-transformer $T$ with $L$ layers, and some restriction $\rho$ such that

$$|\{i \leq n : \rho_n(i) = \ast\}| \geq Cn$$

($C \in (0, 1])$ for all sufficiently large $n$. Choose any $C' < C$.

Then there is a restriction $\rho' \succ \rho$ such that

$$|\{i \leq n : \rho'_n(i) = \ast\}| \geq C'n$$

(1)

(2)

(3)
for all sufficiently large \( n \), and such that there is a \((c \cdot (2^k H + 1))\)-transformer \( T' \) with \( L - 1 \) layers, for some integer \( k \) (depending on \( C \)), where \( H \geq 1 \) is the number of attention heads at each layer and position, such that \( \rho' T = \rho' T' \).

The lemma implies Theorem 1.

Proof of Theorem 1. The output of the transformer is determined by the last activation \( y^{(L)}_n \). Apply the Depth Reduction Lemma iteratively, choosing the constants \( C' \) in the lemma appropriately, until only the zero-th layer remains. Then, after applying the resulting restriction, the final activation \( y^{(L)}_n \) is now computed by \( y^{(0)}_n \), which is determined by a bounded number of input bits.

\[ \square \]

**S1.1 Proving the Depth Reduction Lemma**

In this section, we will prove the Depth Reduction Lemma. We construct the restrictions \( \rho'_n \) separately for each \( n \), on the basis of the given restriction \( \rho_n \). In this process, we will only restrict additional bits, that is, the only case in which \( \rho'_n(i) \) can be different from \( \rho_n(i) \) is that \( \rho'_n(i) \) may be 0 or 1 where \( \rho_n(i) \) was \( \ast \). The construction proceeds in three stages \( \rho^{(1)}_n \), \( \rho^{(2)}_n \), and \( \rho^{(3)}_n = \rho'_n \), which all may restrict additional bits. At the end, we verify that the conclusion of the Depth Reduction Lemma is satisfied for the resulting restriction \( \rho'_n \).

Throughout the proof, we will need a few parameters independent of \( n \): First, we need an integer \( k \) that has to be sufficiently large for the proof to succeed, and will be fixed later in the proof. Second, we need parameters \( \eta \in (0, \frac{1}{2}) \), \( q \in (0, 1) \) and \( \delta > 0 \); they can be chosen as follows:

**Definition 5.** Choose \( \eta \in (0, \frac{1}{2}) \) small, \( q \in (0, 1) \), and \( \delta > 0 \) (such that \((1 + \delta)q \in (0, 1)\)) in such a way as to achieve

\[
(1 - 2\eta) \cdot (1 - (1 + \delta)q) = C'/C
\]

A possible choice to satisfy this is \((1 + \delta)q = \frac{1}{2}, 2\eta = 1 - 2C'/C\).

**Lemma 6 (Stage 1).** There is \( N \) and a restriction \( \rho^{(1)} \succ \rho \) such that

1. each \( \rho^{(1)} \)-free input bit serves as an input to at most \( \frac{1}{\eta} c/C \) many different layer-0 heads, when applying \( \rho^{(1)}_n \).

2. For \( n > N \),

\[
\#\{i \leq n : \rho^{(1)}_n(i) = \ast\} \geq (1 - \eta)Cn
\]

**Proof.** Assume the number of input bits feeding into more than \( \frac{1}{\eta} c/C \) different layer-0 activations is \( \geq \eta Cn \). Then the number of pairs of input bits and depending layer-0 activations is \( \geq \eta Cn \cdot \frac{1}{\eta} c/C = nc \). But there are at most \( nc \) such pairs, because there are \( n \) layer-0 activations, each of which depends on \( \leq c \) inputs. So the number of input bits with \( > \frac{1}{\eta} c/C \) depending layer-0 heads is \( \leq \eta Cn \). We can obtain \( \rho^{(1)}_n \) from \( \rho_n \) by restricting these input bits to some fixed value in \( \{0, 1\} \) (it doesn’t matter which one), and the set \( \{i \leq n : \rho^{(1)}_n(i) = \ast\} \) still has at least \( (1 - \eta)Cn \) elements, for all sufficiently large \( n \). \( \square \)

We write \((h, i)\) for a layer-1 attention head \( h \) \((h = 1, \ldots, H)\) at position \( i \) \((i = 1, \ldots, n)\). Let \( V_\rho(i) \) denote the possible values of \( y^{(0)}_i \). As \( y^{(0)}_i \) depends on \( \leq c \) input bits, we have:

\[
|V_\rho(i)| \leq 2^c
\]
Lemma 15. Let \( \rho \) be a restriction, and \( k \in \mathbb{N} \). Assume the layer-0 head at position \( j \) has more than \( 2^c kH \) many \((k, \rho)\)-depending \((k, \rho)\)-unsatisfied tuples \((h, i, z)\). Then there is a restriction \( \rho' \succ \rho \), restricting only \( \leq c \) additional inputs, such that at least \( kH \) many \((k, \rho')\)-unsatisfied tuples \((h, i, z)\) become \((k, \rho')\)-satisfied.
Proof. Let \( \rho \) be a restriction, and \( k \in \mathbb{N} \). Assume the layer-0 head at position \( j \) has more than \( 2^c kH \) many \((k, \rho)\)-depending \((k, \rho)\)-unsatisfied tuples \(((h, i), z)\). For each \((k, \rho)\)-depending \((k, \rho)\)-unsatisfied tuple \(((h, i), z)\), collect the value \( q' \) of \( y_j^{(0)} \) \((q' \in V_\rho(j))\) resulting in \( A_{((h, i), z), j, \rho} \). There are \( 2^c kH \) such tuples, but only \( 2^c \) possible values \( q' \). So one value \( q \) of them must occur \( kH \) times, by the Pigeonhole Principle. Thus, this \( q \in V_\rho(j) \) is such that
\[
f_{1, h}(z, q) = A_{((h, i), z), j, \rho}
\]
for at least \( kH \) many of these \((k, \rho)\)-depending tuples \(((h, i), z)\).

For such a tuple \(((h, i), z)\), \( j \) now blocks attention on any lower-ranked elements of the ranking. The higher-ranked elements of the ranking can only depend on a total of \( \leq ck \) input bits by Lemma [13]

Definition 16 (Sequence of Restrictions). Define a (finite or infinite) sequence of restrictions \( \rho^{(1)} = \sigma_1 \prec \sigma_2 \prec \ldots \) as follows:

1. \( \sigma_1 := \rho^{(1)} \)

2. Let \( \sigma_i \) be given \((i \geq 1)\). If a layer-0 head has more than \( 2^c kH \) many \((k, \sigma_i)\)-depending \((k, \sigma_i)\)-unsatisfied tuples \(((h, i), z)\), fix \( \leq c \) input bits to make \( \geq kH \) tuples satisfied, using the preceding lemma, obtaining \( \sigma_{i+1} \). Otherwise, terminate the procedure.

Lemma 17. There are \( K, N \) such that for all \( k > K, n > N \), this procedure terminates with \( \rho_n' > \rho_n^{(1)} \) such that

1. We have
\[
\#\{i \leq n: \rho_n'(i) = *\} \geq (1 - 2\eta)Cn
\]

2. No layer-0 head has more than \( 2^c kH \) many \((k, \rho')\)-depending \((k, \rho')\)-unsatisfied tuples \(((h, i), z)\).

Proof. Due to Lemma [10] this procedure can be iterated at most until each tuple \(((h, i), z)\) is \((k, \sigma_i)\)-satisfied, that is, at most
\[
\frac{2^c Hn}{kH} = \frac{2^c n}{k}
\]
times. Let \( U_n \) be the number of times this procedure is iterated \((U_n \leq \frac{2^c n}{k})\). At the end, for \( n > N \),
\[
\#\{i \leq n: (\sigma_U)(i) = *\} \geq (1 - \eta)Cn - cU_n \geq \left(1 - \eta + \frac{2^c c}{k}\right)n
\]

By choosing \( k \) so large that \( \frac{2^c c}{k} \leq \eta C \), we find that
\[
\#\{i \leq n: (\sigma_U)(i) = *\} \geq (1 - 2\eta)Cn
\]
for every \( n > N \). For the second claim, if this were not the case, the procedure would not have terminated at \( \rho_n' \). \( \square \)

Corollary 18 (Stage 2). There is \( K, N \) such that, for each \( k > K \), there is a restriction \( \rho^{(2,k)} > \rho^{(1)} \) such that

1. \( \#\{i \leq n: \rho_n^{(2,k)}(i) = *\} \geq (1 - 2\eta)Cn \) for each \( n > N \)

2. Every \((k, \rho^{(2,k)})\)-unsatisfied \(((h, i), z)\) has at most \( f \leq \frac{2^c}{\eta} C^2 k^2 H/C \) many \((k, \rho^{(2,k)})\)-unsatisfied \((k, \rho^{(2,k)})\)-neighbors.
Proof. Let $\rho^{(2,k)}$ be as given by Lemma 17. The first assertion is immediate from that lemma. For the second assertion, by that lemma, each layer-0 head has at most $2^ckH$ many $(k, \rho^{(2)})$-depending $(k, \rho^{(2)})$-unsatisfied tuples $((h, i), z)$. Using Lemmas 6 and Lemma 13 each input bit has at most $2^{ckH}/n$ many $(k, \rho^{(2)})$-depending $(k, \rho^{(2)})$-unsatisfied tuples. On the other hand, a tuple $((h, i), z)$ can $(k, \rho^{(2)})$-depend on $\leq kc$ inputs by Lemma 13. Multiplying these two bounds gives $\leq 2^{ck}k^2H/C$. □

In order to construct the third and final restriction $\rho^{(3)}_n$, we apply the “probabilistic method”: We define a probability distribution over restrictions $\rho^{(3)}_n$, and show that the probability assigned to restrictions of the type we require is strictly greater than zero, showing that such a restriction exists.

**Definition 19.** Let $k > K$. For each input length $n$, define the distribution over restrictions $\rho^{(3,k)}(K) \succ \rho^{(2,k)}(K)$ that independently assigns to each input position $i \in \{1, \ldots, n\}$ the symbol 1 or 0 with probability $q/2$ each ($q \in (0, 1)$ from Definition 5), and * with probability $1 - q$. On those input bits where $\rho^{(2,k)}(K)(i) \neq *$, we restrict this random restriction to agree with $\rho^{(2,k)}(K)(i)$.

**Definition 20.** Let $k > K$, and consider a $(k, \rho^{(2,k)})$-unsatisfied tuple $((h, i), z)$. By Lemma 9 the sequence

$$\left( y_j^{(0)}_{j, k, i} : s = 1, \ldots, L \right)$$

has length at least $\geq k$.

Define $X^{(3)}_{i,h,k}$ to be the event that, for this tuple, none of the $k$ layer-0 head it depends on $(s = 1, \ldots, k)$ is fixed by $\rho^{(3,k)}$ to the value

$$\operatorname{arg}_{q \in \rho^{(2,k)}(j, k, i, \rho^{(2)})} \max f^{\operatorname{att}}_{1,h}(z, q)$$

(or any element of the argmax, if multiple values achieve this attention weight).

Define $X_{0,k}$ to be the event that more than $(1 + \delta)q$ of the input bits that $\rho^{(2,k)}(K)$ maps to * are set to 0/1 by $\rho^{(3,k)}(K)$ (where $\delta \in (0, 1)$ was fixed in Definition 5).

Our goal will be to show that a nonzero amount of probability mass is assigned to restrictions $\rho'_n$ avoiding all events. We start by individually bounding the probability of each of these events.

**Lemma 21 (X0,k is unlikely).** For any $n > N, k > K$:

$$\mathbb{P}(X_{0,k}) \leq \exp\left( -\frac{\delta^2 q(1 - 2\eta)Cn}{3} \right)$$

**Proof.** Since $\rho^{(2,k)}(K) \geq (1 - 2\eta)Cn$ unrestricted input bits for $n > N$, this follows by a Chernoff bound (Mitzenmacher and Upfal 2017, Theorem 4.4). □

Second, we show that the probability of $X^{(3)}_{i,h,k}$ $(i = 1, 2, \ldots, n, h = 1, \ldots, H)$ decays exponentially in $k$.

**Lemma 22 (X^{(3)}_{i,h,k} is unlikely).** If $((h, i), z)$ is $(k, \rho)$-unsatisfied, then

$$\mathbb{P}(X^{(3)}_{i,h,k}) \leq (1 - (q/2)^c)^k \frac{k^2}{\eta^{2k^c}}$$

for each $i = 1, 2, \ldots, n$ and $h = 1, \ldots, H$. 

5
Proof. Let $Y_{t,i,h,z,k} (t = 1, \ldots, k)$ be the event that the layer-0 activation $y_{j(h,i,z,\rho^{(2)})}^{(0)}$ is not fixed by $\rho^{(3,k)}$ to

$$\arg_{q \in \mathcal{V}_{\rho^{(2,k)}(j(h,i,z,\rho^{(2)})},} \max f_{j,h}^\text{att}(z,q)$$

(20)

Note that

$$X_{i,h}^z = \bigcap_{t=1}^k Y_{t,i}$$

(21)

We have

$$\mathbb{P}(Y_{t,i}^z) \leq 1 - \left(\frac{q}{2}\right)^c \in (0, 1)$$

(22)

Any $Y_{t,i,h,z,k}$ can be statistically dependent on at most

$$c \cdot \frac{1}{\eta} \frac{c}{C} = \frac{1}{\eta} \frac{c^2}{C}$$

(23)

other events $Y_{t',i,h,z,k}$, because each $\rho^{(2,k)}$-free input bit serves as an input to at most

$$\frac{1}{\eta} \frac{c}{C}$$

(24)

layer-0 heads (Lemma 6). Therefore, there is a set of

$$\geq \frac{k}{\frac{1}{\eta} \frac{c^2}{C}}$$

(25)

independent events among these. Call these $Y_{i,h}^1, \ldots, Y_{i,h}^k$. Then

$$X_{i,h}^z \subseteq \bigcap_{s=1}^k Y_{i,h}^s$$

(26)

and thus

$$\mathbb{P}(X_{i,h}^z) \leq \prod_{s=1}^k \mathbb{P}(Y_{i,h}^s) \leq \left(1 - \left(\frac{q}{2}\right)^c \right) \frac{k}{\frac{1}{\eta} \frac{c^2}{C}}$$

(27)

for each $i = 1, 2, \ldots, n$ and $h = 1, \ldots, H$.

Lemma 23. There are $N, K$ such that, for each $n > N$, $k > K$, the probability of avoiding all events

$$\{X_{0,k} \cup \{X_{i,h,k}^z : ((h,i),z) \text{ is } (k,\rho^{(2,k)})\text{-unsatisfied}\}$$

(28)

is strictly greater than zero.

Proof. We apply the Lovász Local Lemma [Mitzenmacher and Upfal 2017 Theorem 6.17]. Each event $X_{i,h,k}^z$ is statistically independent of the set

$$\left\{X_{(j,h',k)}^z : (k,\rho^{(2,k)})\text{-unsatisfied tuples } (j,h',z') \text{ and } (i,h,z) \text{ are not } (k,\rho^{(2,k)})\text{-neighbors}\right\}$$

(29)
The complement of this set has cardinality
\[ \leq f = \frac{2^2 c}{\eta} \pi^2 c H / C \] as concluded in Corollary 18. Set \( A := \frac{1}{k^2}, B := \frac{1}{k^2} \). The number of events \( X_{i,j}^{\pi} \) is bounded by \( 2^f H n \). By the Lovász Local Lemma, it is sufficient to show the following:

\[ \mathbb{P}(X_{i,j}^{\pi}) \leq A(1 - B)(1 - A)^f \quad (31) \]
\[ \mathbb{P}(X_0) \leq B(1 - A)^{2^f H n} \quad (32) \]

The Lovász Local Lemma then guarantees that there is some input restriction \( \rho_{n}^{(3)} \) that avoids all events \( \{X_0 \} \cup \{X_{i,j}^{\pi} : i, h, z \} \). For (31), we need

\[ D \leq A^{1/k}(1 - B)^{1/k}(1 - A)^{f/k} \quad (33) \]

where \( D = (1 - (q/2)^c)^{1/\pi^2 c} \in (0, 1) \). For the first term on the right,

\[ \lim_{k \to \infty} A^{1/k} = \lim_{k \to \infty} \exp \left( -\log(k^2)/k \right) = 1 \]

Also, \( \lim_{k \to \infty} (1 - A)^{f/k} \) equals

\[ \lim_{k \to \infty} \left( 1 - \frac{1}{k^2} \right)^{\frac{2^2 c}{\pi^2 c^2 H / C}} = \lim_{k \to \infty} \left( 1 - \frac{E^2}{k^2} \right)^k = 1 \]

for \( E := \frac{2^2 c}{\pi^2 c^2 H / C} \). So, if we choose \( k \) large enough (independently of \( n \)), the RHS of (33) can be made arbitrarily close to 1, in particular, greater than \( D \). In order to also satisfy (32), we need

\[ \exp \left( -\delta^2 q(1 - 2\eta)C/3 \right) \leq B^{1/n}(1 - A)^{2^f H} \]

which holds for \( n, k \) large enough (again, choosing \( k \) independent of \( n \)). \( \square \)

**Corollary 24.** There are \( K, N \) such that for \( n > N, k > K \), for any \( \rho_{n}^{(3,k)} \) provided by Lemma 23, we have

\[ \{|i \leq n : \rho_{n}^{(3,k)}(i) = \ast\} \geq C' n \]

**Proof.** We have

\[ \{|i \leq n : \rho_{n}^{(3,k)}(i) = \ast\} \geq (1 - 2\eta) \cdot (1 - (1 + \delta)q)C n \]

for all sufficiently large \( n \). The claim follows from the choices in Definition 5. \( \square \)

**Proof of the Depth Reduction Lemma.** After applying \( \rho_{n}^{(3,k)} \), every layer-1 head \( b_{j,i,h} \) depends at most on

1. the \( c \) input bits feeding into \( y_{j}^{(0)} \), and
2. for each \( h = 1, \ldots, H, z \in V_{n}^{(3,k)}(j) \subseteq V_{n}^{(3,k)}(j) \) such that \( (h, j, z) \) is \( (k, \rho_{n}^{(3,k)}) \)-satisfied, at most \( \leq ck \) input bits by the definition of “satisfied”. 7
3. for each \( h = 1, \ldots, H \), \( z \in V_{\rho(3)}(j) \subseteq V_{\rho(2)}(j) \) such that \(((h, j), z)\) is \((k, \rho(2,k))-\text{unsatisfied}\), the input bits that the tuple \( k\)-depends on, of which there are at most \( \leq ck \) by Lemma 13. (Stated differently, every tuple is \((k, \rho(3,k))-\text{satisfied}\).)

Thus, each layer-1 activation \( y^{(1)}_j \) only depends on \( \leq c \cdot (2^k H + 1) \) input bits.

We can thus remove layer 0, convert layer-1 activations \( y^{(1)}_j \) into layer-0 activations \( y^{(0)}_j \), and obtain a \((c \cdot (2^k H + 1))-\text{-transformer performing the same computation as before when } \rho^{(3)} \text{ is applied.} \)

\[ \square \]

### S2  Missing Detail in Soft Attention Proof

In the proof of Lemma 5 on Page 11, the inequality at the end of the first column has the form

\[
\|b - b'\| < \sum a_w \|y_w - y'_w\| \quad (34)
\]

A term is missing: the RHS should be of the form

\[
\|b - b'\| < \sum a_w \|y_w - y'_w\| + \sum |a_w - a'_w| y'_w \quad (35)
\]

The missing term is also small under the assumptions used in the paper.

First, \( y'_w \) is bounded because \( f^{\text{att}} \) and \( f^{\text{act}} \) are Lipschitz functions, and the positional embeddings are assumed to be bounded. These assumptions are used in the \( k=0 \) step of the proof of Lemma 5, and they are necessary for the proof to work.

Second, \( \sum |a_w - a'_w| \) is also in \( O(1/n) \). The next page contains a calculation for this claim.
We want to show that
\[ \sum_{u \neq i} |\hat{a}_{j,u}^{k,h} - \hat{a}_{j,u}^{k,h'}| = O(1/n) \] (1)

To show this, we show that each term is \( O\left(\frac{1}{n^2}\right) \).

First, note \( \hat{a}_{j,u}^{k,h} \in \left[\exp\left(-\frac{2A}{n-1}\right), \exp\left(\frac{2A}{n-1}\right)\right] \) (the upper bound is given in the paper, the lower bound is analogous).

Also, for the unnormalized attention weights, \( |\hat{a}_{j,u}^{k,h} - \hat{a}_{j,u}^{k,h'}| \leq \frac{Q}{n} \) for some constant \( Q \) depending on the parameter matrices and Lipschitz constant of \( f_{\text{att}} \).

Let's fix all indices but \( u \), and write
\[ c_u := \exp(a_u) \in \left[\exp(-A), \exp(A)\right] \] (2)
\[ d_u := \exp(a_u) - \exp(a_u') \] (3)

Because \( |\hat{a}_{j,u}^{k,h} - \hat{a}_{j,u}^{k,h'}| \leq \frac{Q}{n} \), \( a_u \) is bounded, and \( \exp(\cdot) \) is continuous, therefore \( |d_u| \in O\left(\frac{1}{n}\right) \).

Then
\[ \hat{a}_u - \hat{a}_u = \frac{c_u}{\sum y c_y} - \frac{c_u + d_u}{\sum y c_y + d_y} = \frac{c_u(\sum y c_y + d_y) - (c_u + d_u)\sum y c_y}{\sum y c_y(\sum y c_y + d_y)} = \frac{c_u\sum y d_y - d_u\sum y c_y}{\sum y c_y(\sum y c_y + d_y)} \] (4)
\[ \leq \frac{c_u\sum y |d_y| + \frac{C}{n}\sum y c_y}{(\sum y c_y)^2} \leq \frac{\exp(A)C + \frac{C}{n}\sum y c_y}{(\sum y c_y)^2} \] (5)

(for some constant \( C \)). Considering that \( c_u \geq \exp(-A) \), therefore \( \sum y c_y \geq n \exp(-A) \), and this is bounded as
\[ \leq \frac{\exp(A)C + \frac{C}{n}\exp(A)}{n^2\exp(-2A)} = O\left(\frac{1}{n^2}\right) \] (6)
Acknowledgments

Thanks for Gail Weiss, Will Merrill, Navin Goyal, and Satwik Bhattamishra for helpful discussion about the original paper.

References