

Overcoming a Theoretical Limitation of Self-Attention

David Chiang and Peter Cholak

University of Notre Dame
{dchiang, cholak}@nd.edu

Abstract

Although transformers are remarkably effective for many tasks, there are some surprisingly easy-looking regular languages that they struggle with. Hahn shows that for languages where acceptance depends on a single input symbol, a transformer’s classification decisions become less and less confident (that is, with cross-entropy approaching 1 bit per string) as input strings get longer and longer. We examine this limitation using two languages: PARITY, the language of bit strings with an odd number of 1s, and FIRST, the language of bit strings starting with a 1. We demonstrate three ways of overcoming the limitation suggested by Hahn’s lemma. First, we settle an open question by constructing a transformer that recognizes PARITY with perfect accuracy, and similarly for FIRST. Second, we use layer normalization to bring the cross-entropy of both models arbitrarily close to zero. Third, when transformers need to focus on a single position, as for FIRST, we find that they can fail to generalize to longer strings; we offer a simple remedy to this problem that also improves length generalization in machine translation.

1 Introduction

Although transformers (Vaswani et al., 2017) are remarkably effective for many tasks, there are some surprisingly easy-looking formal languages that they struggle with. Hahn (2020) tries to explain some of these by showing (his Lemma 5) that changing a single input symbol only changes the output of a transformer encoder by $O(1/n)$, where n is the input string length. Thus, for a language where acceptance depends on a single input symbol, a transformer might accept or reject strings with perfect accuracy, but for large n , it must do so with low confidence, giving accepted strings a probability just above $\frac{1}{2}$ and rejected strings a probability just below $\frac{1}{2}$. More precisely, as n increases, the cross-entropy approaches its worst possible value of 1 bit per string.

Here, we examine this limitation using two simple regular languages:

$$\text{PARITY} = \{w \in \Sigma^* \mid w \text{ has an odd number of 1s}\}$$
$$\text{FIRST} = \{w \in \Sigma^* \mid w_1 = 1\}$$

where (here and throughout the paper) $\Sigma = \{0, 1\}$. Hahn’s lemma applies to PARITY because the network must attend to all the symbols of the string, and a change in any one of them changes the correct answer. We have chosen FIRST as one of the simplest examples of a language that the lemma applies to. It only requires attention on the first symbol, but the lemma still applies because a change in this symbol changes the correct answer.

Although the lemma might be interpreted as limiting the ability of transformers to recognize these languages, we show three ways that this limitation can be overcome.

First, we show by explicit constructions that transformers do in fact exist that can recognize both languages with perfect accuracy for arbitrary lengths. We have implemented these constructions and verified them experimentally (§3).

As predicted by Hahn’s lemma, our constructed transformers have cross-entropy that approaches 1 bit (that is, just barely better than random guessing) as input length increases. But we show that by adding layer normalization, the cross-entropy can be made arbitrarily close to zero, independent of string length (§4).

In practice, we find, like Bhattamishra et al. (2020a), that transformers cannot learn PARITY. Perhaps more surprisingly, when learning FIRST, transformers can have difficulty generalizing from shorter strings to longer strings. Although this is not a logical consequence of Hahn’s lemma, it is a consequence of the behavior that Hahn’s lemma predicts. Fortunately, this problem can be fixed with a simple modification, multiplying attention logits by $\log n$. This modification also improves length generalization in machine translation (§5).

2 Background

2.1 Notation

If ϕ is a true-or-false statement, we write

$$\mathbb{I}[\phi] = \begin{cases} 1 & \text{if } \phi \text{ is true} \\ 0 & \text{otherwise.} \end{cases}$$

For any $m, n > 0$, we write $\mathbf{0}^{m \times n}$ for the $m \times n$ zero matrix and $\mathbf{I}^{n \times n}$ for the $n \times n$ identity matrix.

2.2 Transformers

Following Hahn (2020), we consider transformer encoders with a sigmoid output layer on a single position. Differently from Hahn (2020), but in line with common practice (Devlin et al., 2019), we prepend a token CLS (for “classification”) and use the encoder output at this token’s position for classifying the string.

We use the original definition of transformers (Vaswani et al., 2017), except for positional encodings.

2.2.1 Input layer

The input to the network is a string $w \in \Sigma^*$. Let $n = |w| + 1$, let $w_0 = \text{CLS}$, and let w_i be the i -th symbol of w .

The input layer has a word embedding and positional encodings,

$$\begin{aligned} \text{WE}: \Sigma &\rightarrow \mathbb{R}^d \\ \text{PE}: \mathbb{N} &\rightarrow \mathbb{R}^d \end{aligned}$$

which are used to compute input vectors for $i = 0, \dots, n$:

$$\mathbf{a}^{0,i} = \text{WE}(w_i) + \text{PE}(i).$$

The word embeddings are typically learned, while the positional encodings vary somewhat. Originally (Vaswani et al., 2017), they were fixed and defined in terms of sine and cosine waves, but they can also be learned (Gehring et al., 2017), in which case they are defined only up to some maximum position. Here, we allow ourselves to define PE as an arbitrary function on all positions. It would seem that to remain in the spirit of the original paper, PE should be easy to compute, independent of w , and parallelizable over positions.

2.2.2 Encoder layers

The body of the encoder is a stack of L layers, each of which has a self-attention sublayer followed by a position-wise feedforward sublayer. For $\ell = 1, \dots, L$, layer ℓ is defined as follows, where $h = 1, \dots, H$, and $i = 0, \dots, n$:

$$\begin{aligned} \mathbf{q}^{\ell,h,i} &= \mathbf{W}^{\text{Q},\ell,h} \mathbf{a}^{\ell-1,i} \\ \mathbf{K}^{\ell,h} &= [\mathbf{W}^{\text{K},\ell,h} \mathbf{a}^{\ell-1,0} \quad \dots \quad \mathbf{W}^{\text{K},\ell,h} \mathbf{a}^{\ell-1,n}]^\top \\ \mathbf{V}^{\ell,h} &= [\mathbf{W}^{\text{V},\ell,h} \mathbf{a}^{\ell-1,0} \quad \dots \quad \mathbf{W}^{\text{V},\ell,h} \mathbf{a}^{\ell-1,n}]^\top \\ \mathbf{c}^{\ell,i} &= \text{LN} \left(\sum_{h=1}^H \text{Att}(\mathbf{q}^{\ell,h,i}, \mathbf{K}^{\ell,h}, \mathbf{V}^{\ell,h}) + \mathbf{a}^{\ell-1,i} \right) \\ \mathbf{h}^{\ell,i} &= \max \left(0, \mathbf{W}^{\text{F},\ell,1} \mathbf{c}^{\ell,i} + \mathbf{b}^{\text{F},\ell,1} \right) \\ \mathbf{a}^{\ell,i} &= \text{LN} \left(\mathbf{W}^{\text{F},\ell,2} \mathbf{h}^{\ell,i} + \mathbf{b}^{\text{F},\ell,2} + \mathbf{c}^{\ell,i} \right) \end{aligned}$$

where boldface lowercase letters stand for vectors in \mathbb{R}^d and boldface uppercase letters stand for matrices in $\mathbb{R}^{d \times d}$. The learned parameters of the model are the \mathbf{W} ’s and \mathbf{b} ’s. The function Att is scaled dot-product attention, defined as

$$\begin{aligned} \text{Att}: \mathbb{R}^d \times \mathbb{R}^{(n+1) \times d} \times \mathbb{R}^{(n+1) \times d} &\rightarrow \mathbb{R}^d \\ \text{Att}(\mathbf{q}, \mathbf{K}, \mathbf{V}) &= \mathbf{V}^\top \text{softmax} \frac{\mathbf{K}\mathbf{q}}{\sqrt{d}} \end{aligned}$$

where the result of the softmax, sometimes written as α , is a vector of *attention weights*. The function LN is layer normalization, whose definition we defer to §4.

2.2.3 Output layer

Finally, the network linearly projects the encoding of CLS to a scalar and applies a sigmoid function:

$$y = \sigma(\mathbf{W}^{L+1} \mathbf{a}^{L,0} + \mathbf{b}^{L+1})$$

where $\mathbf{W}^{L+1} \in \mathbb{R}^{1 \times d}$ and $\mathbf{b}^{L+1} \in \mathbb{R}^{1 \times 1}$. The network accepts w iff the output probability is greater than $\frac{1}{2}$.

3 Exact Solutions

The first way to overcome the limitation suggested by Hahn’s lemma is to show by explicit construction that our two languages can in fact be recognized with perfect accuracy by transformers.

3.1 FFNN for PARITY

Rumelhart et al. (1986) showed that for any n , there is a feedforward neural network (FFNN) that computes PARITY for strings of length exactly n . They

also showed that a randomly initialized FFNN can learn to do this automatically.

Since our construction is partially based on theirs, it may be helpful to review their construction in detail. Let w be the input string, $|w| = n$, and k be the number of 1s in w . The input is a vector \mathbf{x} such that $\mathbf{x}_i = \mathbb{I}[w_i = 1]$. The first layer computes k and compares it against $1, 2, \dots, n$:

$$\mathbf{W}^1 = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} \quad \mathbf{b}^1 = \begin{bmatrix} -0.5 \\ -1.5 \\ \vdots \\ -n + 0.5 \end{bmatrix}$$

so that

$$\mathbf{h}^1 = H(\mathbf{W}^1 \mathbf{x} + \mathbf{b}^1) = \begin{bmatrix} \mathbb{I}[k \geq 1] \\ \mathbb{I}[k \geq 2] \\ \vdots \\ \mathbb{I}[k \geq n] \end{bmatrix}$$

where H is the step function ($H(x) = \mathbb{I}[x > 0]$), applied elementwise.

The second layer adds up the odd elements and subtracts the even elements:

$$\mathbf{W}^2 = [1 \quad -1 \quad \cdots \quad (-1)^{n+1}] \quad \mathbf{b}^2 = -0.5$$

$$y = H(\mathbf{W}^2 \mathbf{h}^1 + \mathbf{b}^2)$$

which is 1 if k is odd and 0 if k is even.

3.2 Transformer for PARITY

Proposition 1. *There is a transformer encoder with sigmoid output layer that recognizes (in the above sense) the language PARITY for strings of arbitrary length.*

Initially, we will construct a transformer encoder without layer normalization (that is, $\text{LN}(\mathbf{x}) = \mathbf{x}$); then we will show how to add layer normalization (§4). Let k be the number of occurrences of 1 in w . All vectors computed by the network have $d = 9$ dimensions; if we show fewer dimensions, assume the remaining dimensions to be zero.

The word and position embeddings are:

$$\text{WE}(\emptyset) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{WE}(1) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\text{WE}(\text{CLS}) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \quad \text{PE}(i) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \frac{i}{n} \\ \cos i\pi \end{bmatrix}.$$

Since we are numbering positions starting from 0, dimension 4 ranges from 0 to $\frac{n-1}{n}$, and dimension 5 is $+1$ for even positions and -1 for odd positions.

We argue that dimension 5, being a cosine wave, is a fairly standard choice, although its period (2) is shorter than the shortest period in standard sinusoidal encodings (2π). Dimension 4 is admittedly not standard; however, we argue that it is a reasonable encoding, and extremely easy to compute.

Thus, the encoding of word w_i is:

$$\mathbf{a}^{0,i} = \begin{bmatrix} \mathbb{I}[w_i = \emptyset] \\ \mathbb{I}[w_i = 1] \\ \mathbb{I}[w_i = \text{CLS}] \\ \frac{i}{n} \\ \cos i\pi \end{bmatrix}.$$

The network has $L = 2$ layers and $H = 2$ heads. The first self-attention layer has one head which finds k , the number of 1s. More precisely, because attention always averages, it must compute the ‘‘average’’ number of 1s, that is, $\frac{k}{n}$, and stores it in dimension 6. It also stores $\frac{1}{n}$ in dimension 7, which we will need later.

$$\mathbf{W}^{\text{Q},1,1} = \mathbf{0}$$

$$\mathbf{W}^{\text{K},1,1} = \mathbf{0}$$

$$\mathbf{W}^{\text{V},1,1} = \begin{bmatrix} \mathbf{0}^{5 \times 5} \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

The second head doesn’t do anything ($\mathbf{W}^{\text{V},1,2} = \mathbf{0}$; the queries and keys can be anything). After the residual connection, we have:

$$\mathbf{c}^{1,i} = \begin{bmatrix} \mathbb{I}[w_i = \emptyset] \\ \mathbb{I}[w_i = 1] \\ \mathbb{I}[w_i = \text{CLS}] \\ \frac{i}{n} \\ \cos i\pi \\ \frac{k}{n} \\ \frac{1}{n} \end{bmatrix}.$$

In the construction of [Rumelhart et al. \(1986\)](#), the next step is to compute $\mathbb{I}[i \leq k]$ for each i , using step activation functions. There are two differences in our construction. First, we have ReLU activation functions, not step activation functions. Second, because attention must sum to one, if n is odd then the even and odd positions will get different attention weights, so the trick of subtracting even positions from odd positions will not work. Instead, we want to compute $\mathbb{I}[i = k]$ (Fig. 1).

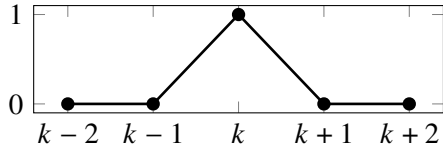


Figure 1: Piecewise linear function equivalent on the integers to $\mathbb{I}[i = k]$.

The first FFNN has two layers. The first is:

$$\mathbf{W}^{\text{F},1,1} = \begin{bmatrix} 0 & 0 & 0 & -1 & 0 & 1 & -1 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 1 \end{bmatrix}$$

$$\mathbf{b}^{\text{F},1,1} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

This gives:

$$\mathbf{h}^{1,i} = \frac{1}{n} \begin{bmatrix} \max(0, k - i - 1) \\ \max(0, k - i) \\ \max(0, k - i + 1) \end{bmatrix}.$$

The second layer linearly combines these three values to get $\mathbb{I}[i = k]$ as desired.

$$\mathbf{W}^{\text{F},1,2} = \begin{bmatrix} \mathbf{0}^{7 \times 3} \\ 1 & -2 & 1 \end{bmatrix} \quad \mathbf{b}^{\text{F},1,2} = \mathbf{0}.$$

After the residual connection, we have:

$$\mathbf{a}^{1,i} = \begin{bmatrix} \mathbb{I}[w_i = \mathbf{0}] \\ \mathbb{I}[w_i = 1] \\ \mathbb{I}[w_i = \text{CLS}] \\ \frac{i}{n} \\ \cos i\pi \\ \frac{k}{n} \\ \frac{1}{n} \\ \frac{\mathbb{I}[i=k]}{n} \end{bmatrix}.$$

The second self-attention layer tests whether position k is even or odd. It does this using two heads, one which attends more strongly to the odd positions, and one which attends more strongly to the even positions; both average dimension 8:

$$\mathbf{W}^{\text{Q},2,1} = [0 \ 0 \ c\sqrt{d} \ 0 \ 0 \ 0 \ 0 \ 0]$$

$$\mathbf{W}^{\text{K},2,1} = [0 \ 0 \ 0 \ 0 \ -1 \ 0 \ 0 \ 0]$$

$$\mathbf{W}^{\text{V},2,1} = \begin{bmatrix} \mathbf{0}^{8 \times 8} \\ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \end{bmatrix}$$

$$\mathbf{W}^{\text{Q},2,2} = [0 \ 0 \ c\sqrt{d} \ 0 \ 0 \ 0 \ 0 \ 0]$$

$$\mathbf{W}^{\text{K},2,2} = [0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0]$$

$$\mathbf{W}^{\text{V},2,2} = \begin{bmatrix} \mathbf{0}^{8 \times 8} \\ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ -1 \end{bmatrix}$$

where $c > 0$ can be any constant. The second FFNN doesn't do anything ($\mathbf{W}^{\text{F},2,1} = \mathbf{b}^{\text{F},2,1} = \mathbf{W}^{\text{F},2,2} = \mathbf{b}^{\text{F},2,2} = \mathbf{0}$). The vector at CLS (position 0) is then

$$\mathbf{a}^{2,0} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 1 \\ \frac{k}{n} \\ \frac{1}{n} \\ \frac{\mathbb{I}[k=0]}{n} \\ s \end{bmatrix}$$

where s has a somewhat complicated value. If n is even, it turns out to be

$$s = (-1)^{k+1} \frac{2 \tanh c}{n^2}$$

which is positive if k is odd and negative if k is even. As predicted by Hahn, it is in $O(1/n)$. If n is odd, the expression for s is more complicated (see Appendix A), but it is still positive iff k is odd, and it is still in $O(1/n)$.

Finally, the output layer is a sigmoid layer that just looks at dimension 9:

$$\mathbf{W}^3 = [0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1] \quad \mathbf{b}^3 = \mathbf{0}$$

$$y = \frac{1}{1 + \exp(-s)}.$$

So the output is greater than $\frac{1}{2}$ iff k is odd.

3.3 Transformer for FIRST

Next, we construct a transformer for FIRST. In line with the common practice of learning per-position word embeddings (Gehring et al., 2017), we use position embeddings that test whether a word is at position 1:

$$\mathbf{a}^{0,i} = \begin{bmatrix} \mathbb{I}[w_i = \mathbf{0}] \\ \mathbb{I}[w_i = 1] \\ \mathbb{I}[w_i = \text{CLS}] \\ \mathbb{I}[i = 1] \end{bmatrix}.$$

The first self-attention layer does nothing ($\mathbf{W}^{\text{V},1,1} = \mathbf{0}$), so after the residual connection, $\mathbf{c}^{1,i} = \mathbf{a}^{0,i}$.

The first FFNN computes a new component (5)

that tests whether $i = 1$ and $w_1 = 1$:

$$\mathbf{W}^{F,1,1} = \begin{bmatrix} -1 & 0 & -1 & 1 \end{bmatrix} \quad \mathbf{b}^{F,1,1} = \mathbf{0}$$

$$\mathbf{W}^{F,1,2} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \quad \mathbf{b}^{F,1,2} = \mathbf{0}$$

$$\mathbf{a}^{1,i} = \begin{bmatrix} \mathbb{I}[w_i = \mathbf{0}] \\ \mathbb{I}[w_i = 1] \\ \mathbb{I}[w_i = \text{CLS}] \\ \mathbb{I}[i = 1] \\ \mathbb{I}[w_i = 1 \wedge i = 1] \end{bmatrix}.$$

(We have chosen $\mathbf{W}^{F,1,1}$ in a slightly unusual way to avoid using the bias term $\mathbf{b}^{F,1,1}$, in anticipation of §4 when we will add layer normalization.)

The second self-attention layer has a single head, which makes CLS focus on position 1.

$$\mathbf{W}^{Q,2,1} = \begin{bmatrix} 0 & 0 & c\sqrt{d} & 0 & 0 \end{bmatrix}$$

$$\mathbf{W}^{K,2,1} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\mathbf{W}^{V,2,1} = \begin{bmatrix} & \mathbf{0}^{5 \times 5} & \\ 0 & 0 & 0 & -\frac{1}{2} & 1 \end{bmatrix}$$

where $c > 0$ is a constant. The second FFNN doesn't do anything ($\mathbf{W}^{F,2,1} = \mathbf{b}^{F,2,1} = \mathbf{W}^{F,2,2} = \mathbf{b}^{F,2,2} = \mathbf{0}$). So at CLS (position 0),

$$\mathbf{a}^{2,0} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ s \end{bmatrix}$$

$$s = \frac{\exp c}{\exp c + n - 1} \left(\mathbb{I}[w_1 = 1] - \frac{1}{2} \right). \quad (1)$$

The final output layer just selects component 6:

$$\mathbf{W}^3 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{b}^3 = \mathbf{0}.$$

So the output probability, $y = \sigma(s)$, is greater than $\frac{1}{2}$ iff $w_1 = 1$. However, it will get closer to $\frac{1}{2}$ as n increases.

3.4 Experiments

We implemented both of the above constructions using modified versions of PyTorch's built-in implementation of transformers (Paszke et al., 2019).¹

¹The code for this and other experiments in this paper are available at <https://github.com/ndnlp/parity>.

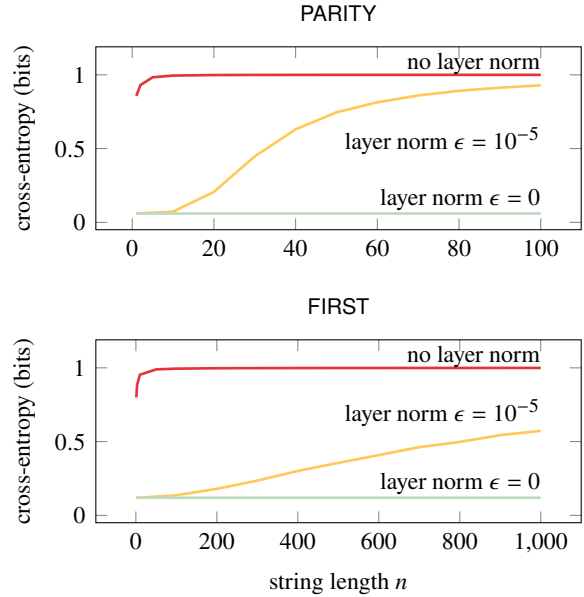


Figure 2: Cross-entropy of exact solutions for PARITY and FIRST computed over 1000 random strings of length n . Without layer norm, the cross-entropy quickly approaches its upper bound of one. With layer norm and $\epsilon > 0$, the cross-entropy is better but still grows with n . With $\epsilon = 0$, cross-entropy is independent of n .

These constructions achieve perfect accuracy for strings with lengths sampled from $[1, 1000]$.

However, in Fig. 2, the red curves (“no layer norm”) show that, as strings grow longer, the cross-entropy approaches its worst possible value of 1 bit per string. We discuss this problem next.

4 Layer Normalization

The second way to mitigate or eliminate the limitation of Hahn’s lemma is layer normalization (Ba et al., 2016), which is defined, for any vector \mathbf{x} , as

$$\text{LN}(\mathbf{x}; \gamma, \beta) = \frac{\mathbf{x} - \text{mean}(\mathbf{x})}{\sqrt{\text{var}(\mathbf{x}) + \epsilon}} \circ \gamma + \beta$$

where the functions mean and var compute the mean and variance, respectively, of the elements of \mathbf{x} , and \circ is the elementwise (Hadamard) product. We fix $\beta = 0$ and $\gamma = 1$, so that the result has approximately zero mean and unit variance. The constant ϵ was not present in the original definition (Ba et al., 2016) but is added in all implementations that we are aware of, for numerical stability.

The original transformer definition performs layer normalization immediately after every residual connection.² In this section, we modify our

²It is also common to place layer normalization before residual connections (Wang et al., 2019; Nguyen and Salazar, 2019), but we follow the original transformer definition here.

two constructions above to use layer normalization. This modification has two steps.

4.1 Removing centering

The first is to nullify the centering effect of layer normalization by making the network compute each value x as well as its negation $-x$. The new word encodings are defined in terms of those in the original construction:

$$\bar{\mathbf{a}}^{0,i} = \begin{bmatrix} \mathbf{a}^{0,i} \\ -\mathbf{a}^{0,i} \end{bmatrix}.$$

Likewise for the self-attention parameters:

$$\begin{aligned} \bar{\mathbf{W}}^{\text{Q},\ell,h} &= \begin{bmatrix} \mathbf{W}^{\text{Q},\ell,h} & \mathbf{0} \end{bmatrix} \\ \bar{\mathbf{W}}^{\text{K},\ell,h} &= \begin{bmatrix} \mathbf{W}^{\text{K},\ell,h} & \mathbf{0} \end{bmatrix} \\ \bar{\mathbf{W}}^{\text{V},\ell,h} &= \begin{bmatrix} \mathbf{W}^{\text{V},\ell,h} & \mathbf{0} \\ -\mathbf{W}^{\text{V},\ell,h} & \mathbf{0} \end{bmatrix}. \end{aligned}$$

Likewise for the position-wise FFNN parameters:

$$\begin{aligned} \bar{\mathbf{W}}^{\text{F},\ell,1} &= \begin{bmatrix} \mathbf{W}^{\text{F},\ell,1} & \mathbf{0} \end{bmatrix} & \bar{\mathbf{b}}^{\text{F},\ell,1} &= \mathbf{b}^{\text{F},\ell,1} \\ \bar{\mathbf{W}}^{\text{F},\ell,2} &= \begin{bmatrix} \mathbf{W}^{\text{F},\ell,2} \\ -\mathbf{W}^{\text{F},\ell,2} \end{bmatrix} & \bar{\mathbf{b}}^{\text{F},\ell,2} &= \begin{bmatrix} \mathbf{b}^{\text{F},\ell,2} \\ -\mathbf{b}^{\text{F},\ell,2} \end{bmatrix}. \end{aligned}$$

Then each layer of activations is

$$\bar{\mathbf{c}}^{\ell,i} = \text{LN} \left(\begin{bmatrix} \mathbf{c}^{\ell,i} \\ -\mathbf{c}^{\ell,i} \end{bmatrix} \right) \quad \bar{\mathbf{a}}^{\ell,i} = \text{LN} \left(\begin{bmatrix} \mathbf{a}^{\ell,i} \\ -\mathbf{a}^{\ell,i} \end{bmatrix} \right).$$

The argument to LN always has zero mean, so that layer normalization does not add or subtract anything. It does scale the activations, but in the case of the two transformers constructed above, any activation layer can be scaled by any positive number without changing the final decisions (see Appendix B).

4.2 Reducing cross-entropy

Furthermore, in any transformer, we can use layer normalization to shrink the cross-entropy as small as we like, contrary to Hahn's Lemma 5. In Hahn's formulation, position-wise functions like layer normalization can be subsumed into his f^{act} , but the lemma assumes that f^{act} is Lipschitz-continuous, and layer normalization with $\epsilon = 0$ is not.

Proposition 2. *For any transformer T with layer normalization ($\epsilon = 0$) that recognizes a language \mathcal{L} , and for any $\eta > 0$, there is a transformer with layer normalization that recognizes \mathcal{L} with cross-entropy at most η .*

Proof. Let d be the number of dimensions in the original vectors of activations, and let L be the number of layers. Then we add a new layer whose self-attention doesn't do anything ($\mathbf{W}^{\text{V},L+1,h} = \mathbf{0}$) and whose FFNN is defined in terms of the original output layer:

$$\begin{aligned} \bar{\mathbf{W}}^{\text{F},L+1,1} &= \begin{bmatrix} \mathbf{I}^d \\ -\mathbf{I}^d \end{bmatrix} & \bar{\mathbf{b}}^{\text{F},L+1,1} &= \begin{bmatrix} \mathbf{0}^d \\ \mathbf{0}^d \end{bmatrix} \\ \bar{\mathbf{W}}^{\text{F},L+1,2} &= \begin{bmatrix} -\mathbf{I}^d & \mathbf{I}^d \end{bmatrix} + \begin{bmatrix} \mathbf{W}^{L+1} & -\mathbf{W}^{L+1} \\ -\mathbf{W}^{L+1} & \mathbf{W}^{L+1} \\ \mathbf{0}^{(d-2) \times d} & \mathbf{0}^{(d-2) \times d} \end{bmatrix} \\ \bar{\mathbf{b}}^{\text{F},L+1,2} &= \begin{bmatrix} \mathbf{b}^{L+1} \\ -\mathbf{b}^{L+1} \\ \mathbf{0}^{d-2} \end{bmatrix}. \end{aligned}$$

This causes the residual connection to zero out all dimensions except two, so that if s was the original output logit, the output of this new layer (before layer normalization) is

$$\bar{\mathbf{a}}^{L+1,i} = \text{LN} \left(\begin{bmatrix} s \\ -s \\ \mathbf{0}^{d-2} \end{bmatrix} \right).$$

Now, if $\epsilon = 0$, layer normalization scales this vector to have unit variance exactly, so it becomes

$$\bar{\mathbf{a}}^{L+1,i} = \begin{bmatrix} \pm\sqrt{d/2} \\ \mp\sqrt{d/2} \\ \mathbf{0}^{d-2} \end{bmatrix}.$$

The new output layer simply selects the first dimension, scaling it by c :

$$\bar{\mathbf{W}}^{L+2} = \begin{bmatrix} c & \mathbf{0} & \mathbf{0}^{d-2} \end{bmatrix} \quad \bar{\mathbf{b}}^{L+2} = 0.$$

Finally, set $c = -\frac{1}{\sqrt{d/2}} \log(\exp \eta - 1)$. If the input string is in \mathcal{L} , then the cross-entropy is $\log \sigma(c\sqrt{d/2}) = \eta$. Similarly, if the input string is not in \mathcal{L} , then the cross-entropy is $\log(1 - \sigma(-c\sqrt{d/2})) = \eta$. \square

However, in practice, ϵ is always set to a nonzero value, which makes layer normalization Lipschitz-continuous, so Hahn's Lemma 5 still applies.

4.3 Experiments

We tested our exact solutions, modified as described above to use layer normalization. Figure 2 shows that layer normalization with $\epsilon > 0$ improves the cross-entropy, but it still grows with n and approaches 1. With $\epsilon = 0$, the cross-entropy is independent of n and, as argued above (Proposition 2), can be made as low as desired.

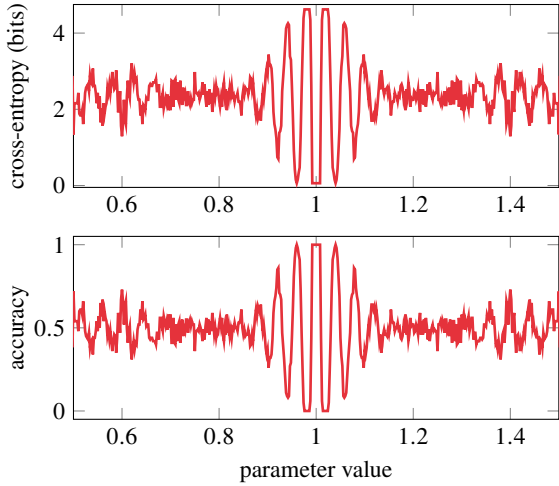


Figure 3: The cross-entropy and accuracy of our solution to PARITY are both extremely sensitive to the parameter $\bar{\mathbf{W}}_{6,2}^{V,1,1}$, which is responsible for computing $\frac{k}{n}$. The correct parameter value is 1.

5 Learnability

In this section, we turn to the question of learnability, which will lead to a third way of overcoming the limitation suggested by Hahn’s lemma.

5.1 Experiments: standard transformers

We tried training transformers on both PARITY and FIRST. Each transformer had the same number of layers and heads and the same fixed positional encodings as the corresponding exact solution. We used $d_{\text{model}} = 16$ for word encodings, self-attention, and FFNN outputs, and $d_{\text{FFNN}} = 64$ for FFNN hidden layers. We used layer normalization ($\epsilon = 10^{-5}$) after residual connections. We used PyTorch’s default initialization and trained using Adam (Kingma and Ba, 2015) with learning rate 3×10^{-4} (Karpathy, 2016). We did not use dropout, as it did not seem to help.

We found, like Bhattamishra et al. (2020a), that a transformer with the above settings was unable to learn PARITY. We tried many other settings as well, to no avail. To give an idea of why our constructed solution, in particular, is difficult to find, Fig. 3 shows the cross-entropy and accuracy of the model if we start with our solution (with layer normalization, $\epsilon = 0$) and vary the parameter $\bar{\mathbf{W}}_{6,2}^{V,1,1}$, which is responsible for computing $\frac{k}{n}$. At 1, it has a cross-entropy of 0 and accuracy of 1, which are both optimal, but the cross-entropy oscillates so rapidly that even a small perturbation of this parameter would make it difficult to recover the solution by gradient descent.

FIRST is much easier to learn, but the bad news is that the learned transformers do not generalize well to longer sentences. Figure 4 (left column) shows that when a transformer is trained from scratch on shorter strings ($n = 10, 30, 100, 300$) and tested on longer strings ($n = 1000$), the accuracy is not perfect. Indeed, for training $n = 10$, the accuracy is hardly better than random guessing.

5.2 Flawed transformer for FIRST

In our solution above (§3.3), the second self-attention layer attended mostly to the first position, but not totally. It relied on the fact that in the second self-attention layer, the values of the non-first positions ($\mathbf{V}_{i,4}^{2,1}$ and $\mathbf{V}_{i,5}^{2,1}$ for $i \neq 1$) are exactly zero and therefore do not contribute to the output.

In practice, because word embeddings are randomly initialized in all dimensions, and are added to every layer via residual connections, it’s unlikely for any activation to be exactly zero. This explains why our exact solution cannot be learned.

But, as a further thought experiment about what the model might be learning instead, consider the following transformer, which uses only a single layer ($L = 1$) and does not zero out the values of the non-first positions. As we will see, it performs worse than the transformer of §3.3 for long strings.

$$\begin{aligned} \mathbf{W}^{Q,1,1} &= \begin{bmatrix} 0 & 0 & c\sqrt{d} & 0 \end{bmatrix} \\ \mathbf{W}^{K,1,1} &= \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix} \\ \mathbf{W}^{V,1,1} &= \begin{bmatrix} \mathbf{0}^{4 \times 4} \\ -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & 0 \end{bmatrix}. \end{aligned}$$

The FFNN doesn’t do anything ($\mathbf{W}^{F,1,1} = \mathbf{b}^{F,1,1} = \mathbf{W}^{F,1,2} = \mathbf{b}^{F,1,2} = \mathbf{0}$), and the final output layer just selects component 5. So if k is the total number of 1s, the final logit at CLS (position 0) would be

$$\begin{aligned} s &= \frac{\exp c - 1}{\exp c + n - 1} \left(\mathbb{I}[w_1 = 1] - \frac{1}{2} \right) \\ &\quad + \frac{1}{\exp c + n - 1} \left(k - \frac{n}{2} \right). \end{aligned}$$

If $c > \log(n - 1)$, then this is positive iff $w_1 = 1$. But if $c \leq \log(n - 1)$, the new second term can be big enough to make the model output an incorrect answer. This suggests that if we train a transformer on strings with length up to N , then the learned parameters will be large enough to classify strings of length up to N correctly, but may misclassify strings longer than N .

This explanation is corroborated by the bottom-left graph in Fig. 4, which shows the attention

weight on the first position of the test string (summed over layers, averaged over strings) as a function of training epoch (starting from random initial parameters). The training strings have varying length (n) and the test strings have fixed length (1000). We might hope that the attention weight would converge to the same value independent of n . But the lower n is, the more the attention weight is diluted, making it easier for the value in position 1 to be outweighed by values in other positions.

5.3 Log-length scaled attention

Fortunately, this problem is easy to fix by scaling the logits of each attention layer by $\log n$, that is, redefining attention as

$$\text{Att}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \mathbf{V}^\top \text{softmax} \frac{\log n}{\sqrt{d}} \mathbf{K} \mathbf{q}. \quad (2)$$

Then taking the model in §5.2 with $c = 1$ gives

$$s = \frac{n-1}{2n-1} \left(\mathbb{I}[w_1 = 1] - \frac{1}{2} \right) + \frac{1}{2n-1} \left(k - \frac{n}{2} \right)$$

which is positive iff $w_1 = 1$. Moreover, scaling is another way to make the cross-entropy low:

Proposition 3. *For any $\eta > 0$ there is a transformer with attention defined as in Eq. (2), and with or without layer normalization, that recognizes FIRST with cross-entropy at most η .*

Proof. Without layer normalization, we can take the model in §3.3, set $c = 1$ and log-scale the attention weights, which changes s from Eq. (1) to

$$s = \frac{n}{2n-1} \left(\mathbb{I}[w_1 = 1] - \frac{1}{2} \right) \\ \frac{1}{4} < |s| \leq \frac{1}{2}.$$

With layer normalization (and $\epsilon > 0$), we can apply the modification of §4 to nullify the centering effect of layer normalization. Then since the variance of $\mathbf{a}^{2,0}$ is $\frac{1}{6}(1+s^2)$, the layer-normalized final logit is

$$\bar{s} = s \left(\frac{1}{6}(1+s^2) + \epsilon \right)^{-\frac{1}{2}}$$

and since $|s| > \frac{1}{4}$,

$$|\bar{s}| > \frac{1}{4} \left(\frac{5}{24} + \epsilon \right)^{-\frac{1}{2}}.$$

In either case, since the final logit has a lower bound not dependent on n , the output layer weights can be scaled as in the proof of Proposition 2 to make the cross-entropy at most η . \square

| | train all test all | train short test long |
|--------------|-----------------------|--------------------------|
| train tokens | 3M+3M | 1M+1M |
| test tokens | 32k+34k | 24k+25k |
| baseline | 32.6 | 11.4 |
| scaled | 32.5 | 12.4 |

Table 1: When training and testing on data with the same length distribution, scaling attention logits has no significant effect on BLEU, but when training on short sentences (≤ 20 tokens) and testing on long sentences (> 20 tokens), scaling helps significantly ($p < 0.01$).

5.4 Experiments: scaled attention

Figure 4 (right column) shows the training of transformers with scaling of attention logits by $\log n$. For all training lengths n , the model is able to learn with perfect test cross-entropy and accuracy.

We see a similar effect on low-resource English-to-Vietnamese machine translation (Table 1), using Witwicky, an open-source implementation of transformers.³ We use all default settings; in particular, residual connections come after layer normalization ($\epsilon = 10^{-5}$). We measure translation accuracy using BLEU (Papineni et al., 2002) and use bootstrap resampling with 1000 samples for significance testing.

When train and test length distributions are the same, scaling attention logits has no significant effect. But if we train only on sentences with median length or shorter (≤ 20 tokens) and test only on sentences longer than median length (> 20 tokens), scaling attention logits by $\log n$ improves BLEU by +1, which is statistically significant ($p < 0.01$).

6 Related Work

Using very different assumptions on the form of transformers and inputs, a number of recent theoretical studies of transformers show that they can solve much more difficult problems than the ones studied here. Transformer encoders can be shown to be universal approximators by fixing the string length and using a number of layers exponential in the length (Yun et al., 2020). Transformer encoder-decoders, where the decoder can run for an unbounded number of steps, have been shown to be Turing-complete (Bhattamishra et al., 2020b; Pérez et al., 2021).

³<https://github.com/tnq177/witwicky>

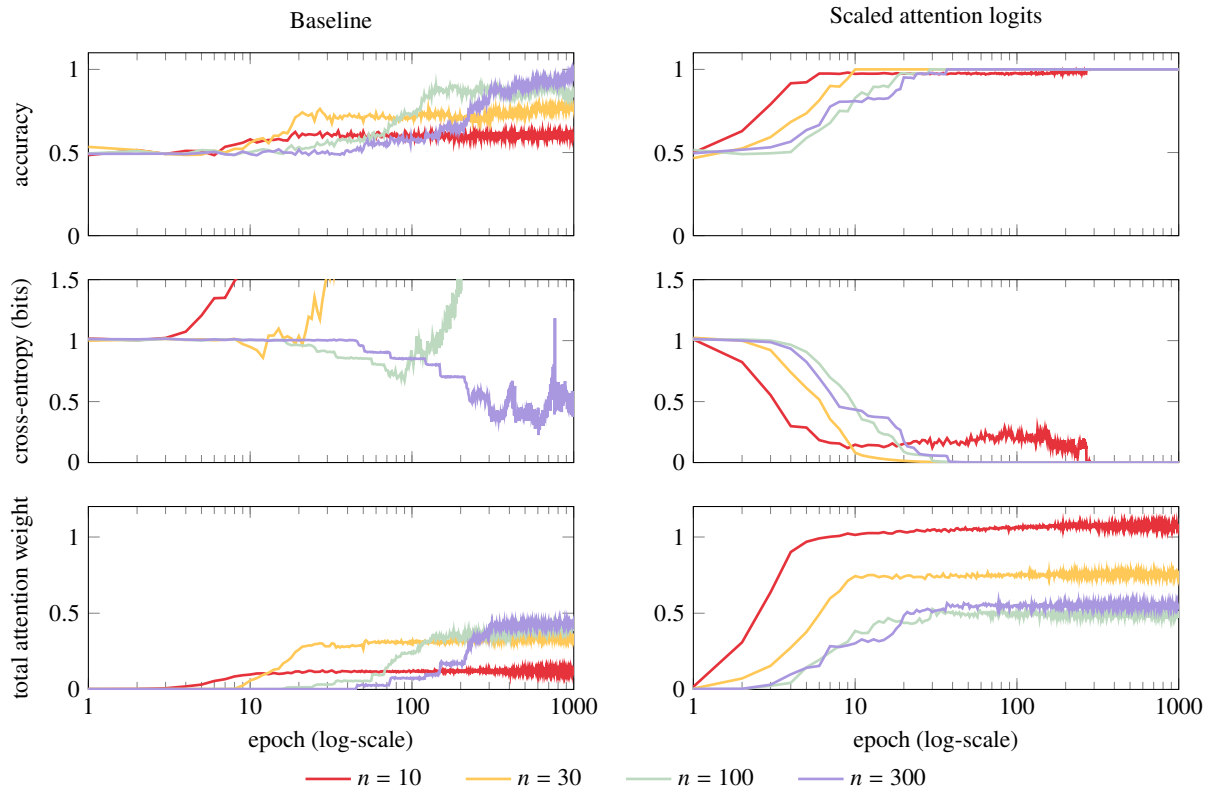


Figure 4: Training a transformer on FIRST. Each epoch has 100 training strings of varying length (see legend) and 100 test strings of length 1000. All curves are averaged over 20 runs. Left: Standard transformer with layer normalization ($\epsilon = 10^{-5}$). Right: Same, with attention logits scaled by $\log n$.

RASP (Weiss et al., 2021) is a simple programming language whose programs can be compiled into transformers. While PARITY can easily be written in RASP, this does not imply in itself the existence of transformers that can recognize PARITY, for two reasons. First, RASP’s aggregate operation (which corresponds to attention) always attends uniformly to a subset of positions, unlike softmax attention. Second, RASP’s elementwise operations are embedded directly in the output transformer; they are not translated into FFNNs.

Bhattachamishra et al. (2020a) carry out theoretical and experimental studies of transformers for various formal languages. The theoretical results are for a different variant of transformers than ours (transformer encoders with self-attention masked so that each position attends only to previous positions), and focus on such transformers’ ability to maintain counters that are constrained to be non-negative. Their experimental results suggest that transformers have difficulty learning some regular languages, including PARITY.

7 Conclusion

We’ve seen that the questions of (a) whether a neural network can recognize a language, (b) whether it can achieve low cross-entropy on a language, and (c) whether it can learn to recognize a language are three separate questions, because we have given examples of (a) without (b) and (b) without (c).

Namely, our explicit construction for PARITY shows that a neural network can recognize a language with perfect accuracy (a) but poor cross-entropy (b). Adding layer normalization ($\epsilon = 0$) enables it to achieve low cross-entropy (b), but still does not learn well (c). We observe that because the answer to (b) can hinge on small details of the model, (b) is not probably not very useful as a way of measuring expressivity.

However, we did find that the limited influence of a single input symbol, implied by Hahn’s lemma, has a serious practical implication for learnability (c). Namely, transformers can fail to generalize from shorter training strings to longer testing strings. Our proposed fix, scaling attention logits by $\log n$, is easy to implement and very effective on a real machine translation task.

Acknowledgements

We would like to thank Toan Nguyen for assistance with his machine translation code, and Gail Weiss for catching some mistakes.

This paper is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract #FA8650-17-C-9116. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. [Layer normalization](#). arXiv:1607.06450.
- Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. 2020a. [On the ability and limitations of Transformers to recognize formal languages](#). In *Proc. EMNLP*, pages 7096–7116.
- Satwik Bhattamishra, Arkil Patel, and Navin Goyal. 2020b. [On the computational power of Transformers and its implications in sequence modeling](#). In *Proc. CoNLL*, pages 455–475.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proc. NAACL HLT*, pages 4171–4186.
- Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. [Convolutional sequence to sequence learning](#). In *Proc. ICML*, pages 1243–1252.
- Michael Hahn. 2020. [Theoretical limitations of self-attention in neural sequence models](#). *Trans. ACL*, 8:156–171.
- Andrej Karpathy. 2016. [3e-4 is the best learning rate for Adam, hands down](#). Twitter.
- Diederik P. Kingma and Jimmy Lei Ba. 2015. [Adam: A method for stochastic optimization](#). In *Proc. ICLR*.
- William Merrill, Vivek Ramanujan, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. [Effects of parameter norm growth during transformer training: Inductive bias from gradient descent](#). In *Proc. EMNLP*, pages 1766–1781.
- Toan Q. Nguyen and Julian Salazar. 2019. [Transformers without tears: Improving the normalization of self-attention](#). In *Proc. International Workshop on Spoken Language Translation*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: a method for automatic evaluation of machine translation](#). In *Proc. ACL*, pages 311–318.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [PyTorch: An imperative style, high-performance deep learning library](#). In *Proc. NeurIPS*.
- Jorge Pérez, Pablo Barceló, and Javier Marinkovic. 2021. [Attention is Turing-complete](#). *Journal of Machine Learning Research*, 22(75):1–35.
- D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1986. *Learning Internal Representations by Error Propagation*, pages 318–362. MIT Press, Cambridge, MA, USA.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Proc. NeurIPS*, pages 5998–6008.
- Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019. [Learning deep Transformer models for machine translation](#). In *Proc. ACL*, pages 1810–1822.
- Gail Weiss, Yoav Goldberg, and Eran Yahav. 2021. [Thinking like Transformers](#). In *Proc. ICML*.
- Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. 2020. [Are Transformers universal approximators of sequence-to-sequence functions?](#) In *Proc. ICLR*.

A Correctness of PARITY Construction

In §3.2, we constructed a transformer that recognizes PARITY; here we fill in details of calculating $s = \mathbf{a}_9^{2,0}$. If n is even, the first head computes

$$\begin{aligned} \mathbf{q}^{2,1,0} &= c\sqrt{d} \\ \mathbf{K}_{i,1}^{2,1,0} &= -\cos i\pi = (-1)^{i+1} \\ \alpha_i^{2,1,0} &= \frac{\exp(-1)^{i+1}c}{\frac{n}{2}(\exp c + \exp -c)} \\ \mathbf{V}_{i,9}^{2,1,0} &= \frac{\mathbb{I}[i = k]}{n}. \end{aligned}$$

Similarly, the second head computes

$$\begin{aligned} \mathbf{q}^{2,2,0} &= c\sqrt{d} \\ \mathbf{K}_{i,1}^{2,2,0} &= \cos i\pi = (-1)^i \\ \alpha_i^{2,2,0} &= \frac{\exp(-1)^i c}{\frac{n}{2}(\exp c + \exp -c)} \\ \mathbf{V}_{i,9}^{2,2,0} &= -\frac{\mathbb{I}[i = k]}{n}. \end{aligned}$$

Then

$$\begin{aligned} s = \mathbf{a}_9^{2,0} &= \frac{1}{n}\alpha_k^{2,1,0} - \frac{1}{n}\alpha_k^{2,2,0} \\ &= \frac{\exp(-1)^{k+1}c - \exp(-1)^k c}{\frac{n^2}{2}(\exp c + \exp -c)} \\ &= (-1)^{k+1} \frac{\exp c - \exp -c}{\frac{n^2}{2}(\exp c + \exp -c)} \\ &= (-1)^{k+1} \frac{2 \tanh c}{n^2} \end{aligned}$$

is negative if k is even and positive if k is odd.

If n is odd, calculating s is more complicated because there are unequal numbers of more- and less-attended positions. The attention weights are

$$\begin{aligned} \alpha_i^{2,1,0} &= \frac{\exp(-1)^{i+1}c}{\underbrace{\frac{n-1}{2}\exp c + \frac{n+1}{2}\exp -c}_{Z_1}} \\ \alpha_i^{2,2,0} &= \frac{\exp(-1)^i c}{\underbrace{\frac{n+1}{2}\exp c + \frac{n-1}{2}\exp -c}_{Z_2}} \\ s &= \frac{(\exp(-1)^{k+1}c)Z_2 - (\exp(-1)^k c)Z_1}{nZ_1Z_2}. \end{aligned}$$

If k is even,

$$\begin{aligned} s &= \frac{\frac{n-1}{2}\exp -2c - \frac{n-1}{2}\exp 2c}{nZ_1Z_2} \\ &= -\frac{(n-1)\sinh 2c}{nZ_1Z_2} < 0 \end{aligned}$$

whereas if k is odd,

$$\begin{aligned} s &= \frac{\frac{n+1}{2}\exp 2c - \frac{n+1}{2}\exp -2c}{nZ_1Z_2} \\ &= \frac{(n+1)\cosh 2c}{nZ_1Z_2} > 0. \end{aligned}$$

B Scale-Invariance of PARITY and FIRST Constructions

In §4.1, we claimed that the scaling effect of layer normalization has no effect on the decisions of our constructions for PARITY and FIRST. This is related to the property of approximate homogeneity studied by Merrill et al. (2021).

In general, we rely on the fact that the FFNNs we use all have no bias terms ($\mathbf{b}^{F,\ell,1}$ and $\mathbf{b}^{F,\ell,2}$), so the FFNNs are 1-homogenous (scaling the input scales the output by the same amount). For the self-attentions, our $\mathbf{W}^{Q,\ell,h}$ all have a constant factor c built into them, so any scaling of the input can be absorbed into this constant.

For PARITY, suppose that layer normalization scales \mathbf{c}^ℓ by C_ℓ and \mathbf{a}^ℓ by A_ℓ .

$$\bar{\mathbf{c}}^{1,i} = C_1 \begin{bmatrix} \mathbf{c}^{1,i} \\ -\mathbf{c}^{1,i} \end{bmatrix}$$

Because the first FFNN has no bias term,

$$\bar{\mathbf{a}}^{1,i} = A_1 C_1 \begin{bmatrix} \mathbf{a}^{1,i} \\ -\mathbf{a}^{1,i} \end{bmatrix}$$

In the second self-attention layer, the attention logits and the values are scaled by $A_1 C_1$. We're only interested in what happens to $s = \mathbf{c}_9^{2,0}$. If n is even, s becomes:

$$\bar{s} = (-1)^{k+1} \frac{2C_2 A_1 C_1 \tanh A_1 C_1 c}{n^2}.$$

Since the second FFNN is the identity function, its layer normalization has no effect ($A_2 = 1$). So the final logit is \bar{s} , which is still negative if k is even and positive if k is odd. Similarly if n is odd.

For FIRST, again suppose that layer normalization scales \mathbf{c}^ℓ by C_ℓ and \mathbf{a}^ℓ by A_ℓ . As before,

$$\bar{\mathbf{a}}^{1,i} = A_1 C_1 \begin{bmatrix} \mathbf{a}^{1,i} \\ -\mathbf{a}^{1,i} \end{bmatrix}$$

In the second self-attention layer, the attention logits and the values are scaled by $A_1 C_1$. We're only interested in what happens to $s = \mathbf{c}_6^{2,0}$:

$$\bar{s} = \frac{\exp A_1 C_1 c}{\exp A_1 C_1 c + n - 1} C_2 A_1 C_1 \left(\mathbb{I}[w_1 = 1] - \frac{1}{2} \right)$$

Since the second FFNN is the identity function, $A_2 = 1$. So the final logit is \bar{s} , which is still positive if $w_1 = 1$ and negative otherwise.