ATTENTION IS OPTIMAL BUILDING BLOCK

Dan Navon

November 25, 2021

1 Basic Idea

1.1 Introducing

As shown in [1] if we throw all the activations from the transformers we can remain with polynomial of degree 3^L where L is the depth that is good approximation for the transformer in the sense that the induced networks achieves similar performance up to small degradation, hence we can think on this linear transformer for our analysis and make experiments at the end for standard transformers.

And again by [1] the expressive power of this transformer as measured by the separation rank, is as long as $L > \log_3 d_x$

$$sep_{(A,B)}\left(y\right) = \Theta\left(3^{L\cdot d_x}\right) \tag{1}$$

where d_x is the layer width, and L is the depth, and the 3 is since each layer y_i in the linear transformer is polynomial of degree 3, and now let ask ourselves what if we change the degree of each specific layer would we be able to achieve better expressive power with the same budget.

1.2 Formalizing

Let y_i^d denote layer of linear transformer but with degree d i.e

$$y_i^d\left(X\right) := X\prod_{k=1}^{d-1}\left(W_k X^t\right) \tag{2}$$

then if Y_L^k is linear transformer with p layers each from degree d then the obtained expressive power is

$$sep_{(A,B)}\left(Y_L^d\right) = \Theta\left(d^{L\cdot d_x}\right) \tag{3}$$

while the number of parameters we used is

$$N_{L,d} = L\cdot d\cdot N_{d_x} \tag{4}$$

as long as $L > \log_d d_x$ where N_{d_x} is the number of parameters of some specific weights matrix W_k, hence for fixed budget of parameters B we are looking after

$$\arg\max_{d,L} d^{L\cdot d_x}$$

$$L\cdot d = \frac{B}{N_{d_x}}$$

or alternatively we are looking to maximize

$$f\left(d\right) := d^{\frac{B\cdot d_x}{dN_{d_x}}} = \left(d^{\frac{1}{d}}\right)^{\frac{B\cdot d_x}{N_{d_x}}}$$

And observe that

$$d^* := \arg\max_d \left(d^{\frac{1}{d}}\right)^{\frac{B\cdot d_x}{N_{d_x}}} \underbrace{=}_{\frac{B\cdot d_x}{N_{d_x}}>0} \arg\max_d d^{\frac{1}{d}} = \arg\max_d \frac{\ln d}{d}$$

finally d_c is critical point of $h(d) := \frac{\ln d}{d}$ iff

$$0 = h^{'}(d) = \frac{1}{d^2} - \frac{\ln d}{d^2} \iff \ln d = 1 \iff d = e \tag{5}$$

hence $d_c = e$ is the only critical point of h and since,

$$d > e \implies h^{'}(d) = \frac{1-\ln d}{d^2} < \frac{1-\ln e}{d^2} = 0$$
$$d < e \implies h^{'}(d) = \frac{1-\ln d}{d^2} > \frac{1-\ln e}{d^2} = 0$$

But we know that

$$d > e \implies h(d) = h(e) + \underbrace{\int_e^d h^{'}(s)\, ds}_{<0} < h(e)$$
$$d < e \implies h(d) = h(e) - \underbrace{\int_d^e h^{'}(s)\, ds}_{>0} < h(e)$$

Hence the only maximum is obtained for $d = e \approx 2.71$ and the function is monotone decreasing when we go farther on both sides, hence if we want $d^*_{\mathbb{N}} \in \mathbb{N}$ then it must hold that $d^*_{\mathbb{N}} \in \{2, 3\}$ finally

$$h(3) > h(2) \iff \frac{\ln 3}{3} > \frac{\ln 2}{2} \iff 2\ln 3 > 3\ln 2 \iff \ln 3^2 > \ln 2^3 \iff 3^2 > 2^3 \iff 9 > 8$$

Conclusion 1.1 *Concluding $d^*_{\mathbb{N}} = 3$ while $d^* = e$ and the optimal building block for the transformer mechanism should be polynomial of degree 3 which is exactly the attention mechanism, as long as we assume that it should be polynomial with integer degrees.*
If in someway we know to create building blocks with fractional degrees then further improvement can be achieved by taking each building block to be with degree e.

2 Missing Pieces

Task 1 Extend the proof from [1] into rank k polynomials within each building block.
Task 2 Conduct experiments
Task 3 How to build fractional transformers with $d = e$?
Task 4 It also explain the mixer MLP results (Need Detailed Analysis).

3 Experiments

Since Task-1 is time consuming and low risk, let start with conducting the experiments first.

3.1 Experiment Design

For each $d \in \{2, 3, 4\}$ we will train DNN $\mathcal{T}_d$ that composes from the layers $\left(L^d_1, ..., L^d_p\right)$ i.e

$$\mathcal{T}_d(x) := L^d_p \circ L^d_{p-1} \circ ... \circ L^d_1(x) \tag{6}$$

where $\forall i, \quad L^d_i := \sigma \circ Q^d_i(x)$ and $Q^d_i(x)$ is polynomial of degree d in x, and σ is some activation maybe combined with normalization.
Summarizing we will train several FC-DNN's but in each of them the linear layers would be replaced by polynomials of degree d, and our goal would be to find which d achieves the best results, for fixed parameters budget B, where our theory claims that $d = 3$ is the optimal one.

Comment 3.1 *Our polynomial for degree d is of the form $P_d(x) := \prod_{k=1}^{d}(W_k x)$ In particular it doesn't contains all the coefficients and it is sparse polynomial.*

3.2 Experiment Results

Currently in the middle, and suffers from convergence problem which I'm trying to solve.

4 Proof Generalization

4.1 Strategy

So now our goal is to find the expressive power of the relevant networks, so let $\mathcal{T}_d^p$ be the network with depth p where each layer has degree d, as defined above, then we want to measure the expressive power of the network as measured by the separation rank, then the most obvious way is to start from the existing proof from for the transformer with the attention mechanism and try to extend it into the case where the degree is d instead of 3.

4.2 Proxy Measure

And for start let start with some proxy measure, of taking the degree of the final polynomial as some proxy for the complexity and maybe the expressiveness of the entire polynomial, then indeed,

$$degree\left(\mathcal{T}_p^d\right) = degree\left(L_p \circ ... \circ L_1\right) = degree\left(L_p\right) \times ... \times degree\left(L_1\right) = d^p \tag{7}$$

And since if N_{d_x} is the number of parameters in each specific layer then the total number of parameters in the network is $d \cdot p \cdot N_{d_x} = B$ then fixing the budget we have $p = \frac{B}{N_{d_x}} \cdot \frac{1}{d}$ combining with (8) we have

$$degree\left(\mathcal{T}_p^d\right) = \left(d^{\frac{1}{d}}\right)^{\frac{B}{N_{d_x}}} \tag{8}$$

Hence the total degree is maximized when d is taken to maximize the expression $d^{\frac{1}{d}}$ which results in $d = e$ when $d \in \mathbb{R}$ and $d = 3$ if we require $d \in \mathbb{N}$.

4.3 Separation Rank

Now let try to prove for the separation rank, in the same way like before, then we already have the proof for the case $p = 3$ and instead of attacking immediately the general case let start with the case $d = 2$ first

4.3.1 Warn-Up d=2

In this case we have

$$L_i^d(x) = x^T W_{i1}^T W_{i2} x \tag{9}$$

Or alternatively

$$L_i^d(x) = x^T W x \tag{10}$$

where $rankW = d_x$ i.e it can be factorized into the form $W = W_{i1} W_{i2}^T$ and just to feel how this formula behaves then

$$L_2 \circ L_1(x) = L_2(L_1(x)) = L_2\left(x^T W_1 x\right) = \left(x^T W_1 x\right) W_2 \left(x^T W_1 x\right) = \left(x^T W_1 x\right) W_2 \left(x^T W_1 x\right) \tag{11}$$

and all of this is after we forget the W_O in the end of each layer to normalize the dimensions.

4.3.2 Naive Approach

Then first each element in polynom of the form $q_{k_1,...,k_d}(x_1, ..., x_d) := \prod_{j=1}^{d} x_j^{k_j}$

References

[1] Yoav Levine Noam Wies Daniel Jannai Dan Navon Yedid Hoshen Amnon Shashua. “THE INDUCTIVE BIAS OF IN-CONTEXT LEARNING: RETHINKING PRETRAINING EXAMPLE DESIGN”. In: *CVPR 2021* 0.0 (2021), p. 40. DOI: `https://arxiv.org/pdf/2110.04541.pdf`.