UNIVERSALITY OF SCOT WITH RESPECT TO ANO

DEVANSH TRIPATHI[1]
ETH Zürich

In this draft, I will highlight the differences between the two proofs for universality of scOT. First of all, the proof provided in the "scOT_universality" draft does not aim to convert scOT to an ANO rather it directly shows the existence of a scOT with desired accuracy while I make choices such that SwinV2 gets reduce to ANO.

1. Proof in "scOT_universality" draft

In the draft, window is taken to be whole domain. Q, K matrices are taken to be 0. W^h is set to be $\mathbb{I}_{C\times C}$, $V^h \in \mathbb{R}^{C\times C}$ with $V_{ij} = \delta_{ih}\delta_{jh}$ with B as neural network approximating a function

$$\overline{B}^h(x, y) = \ln(2 + \eta_h(x - y)) - 1, \tag{1.1}$$

where η_h is an enumeration of Fourier basis together with 0 function. This choice of B to approximates eq 1.1 is taken to simplify the expression with exponent. This will result in simplification of $W - MSA$ as average of $v(x)$ over the whole domain plus convolution of $\bar{a}$ with the Fourier basis vector as an extra term.

$$W - MSA(v)(x) = \int_D v(y)dy + \frac{1}{2}\int_D (\eta_1(x - y), \ldots, \eta_H(x - y))^\top \bar{a}(y)dy$$

where η_h are the enumeration of the real Fourier basis together with 0 function and $\bar{a}$ is the average of the input function $a \in C(D; \mathbb{R}^n)$. The aim is to prove

$$\sup_{u\in K} \|\mathcal{S}(u) - \mathcal{S}^*(u)\|_{C^{s'}} \leq \epsilon$$

where $\mathcal{S} : C^s(\overline{D}; \mathbb{R}^n) \to C^{s'}(\overline{D}; \mathbb{R}^n)$ and $\mathcal{S}^* : K \to C^{s'}(\overline{D}; \mathbb{R}^n)$ for $K \subset C^s(\overline{D}; \mathbb{R}^n)$ be a compact subset.

Now, since $\mathcal{S}$ is a continuous operator, there exists an MLP Ψ_h for all h, see [1], mapping Fourier coefficients of $\bar{a}(\cdot + x)$ to an approximation to those of $\mathcal{S}(P_N(\bar{a}))$ where P_N is the projection operator on L^2_N.

$$\Psi_h\left(\int_D \eta_h(y)\bar{a}(y + x)dy\right) = \int_D \eta_h(y)\mathcal{S}(P_N\bar{a}(\cdot - x))(y)dy - (v_1'(x))_h.$$

Then $\mathcal{S}^*$ is constructed as

$$S^*(a)(x) = P_N\mathcal{S}(P_N\overline{a})(x).$$

Then for large enough N, author says error will be smaller than ϵ.

2. Proof with ANO (my proof)

v is the output of the embedding operator and $v(x) \in \mathbb{R}^C$ which is the input for SwinV2 block. First I have tried to simplify windowed multi head self attention operator with the following choices: $H = 1$ (which will convert multi head attention to single head), C is the latent space dimension and take $m = 1$ (so window is just the whole domain).

While, I take a bit different choices to simplify the exponent expression: since $X, Q, K, V \in \mathbb{R}^{m\times C}$, taking $m = 1$ (window is just whole domain) simplifes them as $X, Q, K, V \in \mathbb{R}^{1\times C}$ (row

[1]Seminar für Angewandte Mathematik, HG E 62.2, Rämistrasse 101, 8092 Zürich, Switzerland
devansh.tripathi@sam.math.ethz.ch.

vectors). Also, $V = XW^V$ (linear projection of input sequence of tokens X), I make a choice of $W^V = X^\dagger$ (pseudo-inverse) and Q, K which are defined as $Q = XW^Q, K = XW^K$ can be anything. Hence, when we multiply $v(x) \in \mathbb{R}^C$ with $Q \in \mathbb{R}^{1\times C}$ we get a scalar and same for $Kv(x) \in \mathbb{R}$. I take positional bias B as 0 vector instead of eq 1.1 and W^h as $(XX^\dagger)^{-1}\mathrm{Id}$ where $XX^\dagger \in \mathbb{R}$ (let $XX^\dagger = \alpha \neq 0$) (since X is non-zero vector, α will automatically be non-zero).

With these choices, $\cos(Qv(x), Kv(x))$ becomes 1 (since this is cos of angle between the scalars or say vectors of length 1). These choices simplify the exponenent expression as average of $v(x)$ over the whole domain.

$$W - MSA(v)(x) = \frac{1}{|D|}\int_D v(y)dy.$$

Then I show that SwinV2 block is an ANO with a residual connection (I called it ResANO). This ResANO has the core structure of ANO (nonlinearity and average as nonlocality) hence it is also universal.

2.1. **Proof for scOT.** Since scOT is a composition of SwinV2 blocks, ConvNeXt block with some operations of merging, expansion etc. (all these operations are linear transformations that changes the dimensions) and SwinV2 is an ANO and ConvNeXt is an ANO (shown in "poseidon.pdf"), scOT is universal (or should say can be reduced to ANO).

References

[1] Samuel Lanthaler, Zongyi Li, and Andrew M. Stuart. Nonlocality and nonlinearity implies universality in operator learning, 2024. URL `https://arxiv.org/abs/2304.13221`.