Are Transformers Universal Approximators of Seq2Seq Functions?
Chulhee Yun†, Srinadh Bhojanapalli‡, Ankit Singh Rawat‡, Sashank J. Reddi†, Sanjiv Kumar‡
†MIT, ‡Google Research NYC

Transformers

- Transformers enable state-of-the-art performance in various NLP tasks. E.g., BERT, GPT2, XLNet, ...
- A Transformer block $t^{h,m,r}$ is a map from $\mathbb{R}^{d \times n}$ to $\mathbb{R}^{d \times n}$, defined by two key components:
  - Self-attention layer
    $$\text{Attn}(X) = X + W_o \begin{bmatrix} \text{head}(X) \\ \text{head}_d(X) \end{bmatrix},$$
    where $\text{head}(X) = (W^i_o X) \cdot \text{softmax}[(W^i_q X)^T W^i_k X]$, $W_o \in \mathbb{R}^{d \times bm}$, $W^i_q \in \mathbb{R}^{m \times d}$.
  - Token-wise feed-forward layer
    $$\text{FF}(X) = \text{ReLU}(W_1 \cdot \text{Attn}(X)), \quad W_1 \in \mathbb{R}^{r \times d}, \quad W_2 \in \mathbb{R}^{d \times r'}$$

Our Contributions

- What is the class of sequence-to-sequence functions that Transformers can represent?
- Due to parameter sharing, it is not clear if Transformers can represent arbitrary sequence-to-sequence functions.

We show that a Transformer network with large enough depth is able to approximate any continuous and permutation equivariant function up to arbitrary accuracy.

Definitions

- For a function $f : \mathbb{R}^{d \times n} \rightarrow \mathbb{R}^{d \times n}$ is permutation equivariant if for any permutation matrix $P \in \mathbb{R}^{n \times n}$, $f(X) = f(X)P$.
- For a matrix $A \in \mathbb{R}^{d \times n}$, we denote the entrywise $p$-norm of $A$ as $\|A\|_p$.
- Define function distance $d_p(f, g) := (\int \|f(X) - g(X)\|_p^d dX)^{1/p}$.

Main Theorems

$$\mathcal{F}_{PE} := \{ f : \mathbb{R}^{d \times n} \rightarrow \mathbb{R}^{d \times n} | f \text{ is continuous, permutation equivariant, and has compact support.} \},$$
$$\mathcal{T}^{h,m,r} := \{ g : \mathbb{R}^{d \times n} \rightarrow \mathbb{R}^{d \times n} | g \text{ is a composition of Transformer blocks} t^{h,m,r} \}.$$

**Theorem 1.** A Transformer network with constant width and large enough depth can approximate any function $f \in \mathcal{F}_{PE}$ to arbitrary accuracy; i.e., let $1 \leq p < \infty$ and $\epsilon > 0$, then for any given $f \in \mathcal{F}_{PE}$, there exists a Transformer network $g \in \mathcal{T}^{2,1,4}$ such that $d_p(f, g) \leq \epsilon$.

If the network has positional encoding, one can remove permutation equivariance:
$$\mathcal{F}_{CD} := \{ f : \mathbb{R}^{d \times n} \rightarrow \mathbb{R}^{d \times n} | f \text{ is continuous on a compact domain } \mathbb{D} \subset \mathbb{R}^{d \times n} \},$$
$$\mathcal{T}^{h,m,r}_{PE} := \{ gP(X) = g(X + E) | g \in \mathcal{T}^{h,m,r} \text{ and } E \in \mathbb{R}^{d \times n} \text{ is learnable.} \}$$

**Theorem 2.** Let $1 \leq p < \infty$ and $\epsilon > 0$, then for any given $f \in \mathcal{F}_{CD}$, there exists a Transformer network $g \in \mathcal{T}^{2,1,4}_{PE}$ such that we have $d_p(f, g) \leq \epsilon$.

Contextual Mappings

**Def.** Consider a finite set $L \subseteq \mathbb{R}^{d \times n}$. A map $q : L \rightarrow \mathbb{R}^{1 \times n}$ defines a contextual mapping if the map satisfies the following:
- If any $L \in L$, the $n$ entries in $q(L)$ are all distinct.
- For any $L, L' \in L$, with $L \neq L'$, all entries of $q(L)$ and $q(L')$ are distinct.

Self-attention layers can implement a permutation equivariant contextual mapping!

Proof Sketch

- Approximate $f \in \mathcal{F}_{PE}$ with support $[0, 1]^{d \times n}$ by a piecewise constant function $\mathcal{F}$ on $\delta$-cubes at each $L \in \mathcal{G} = \{0, \delta, 2\delta, \ldots, 1 - \delta\}^{d \times n}$.
- Using FF, quantize $[0, 1]^{d \times n}$ to the grid $\mathcal{G}$.
- Using Attn, implement a perm. equivariant contextual mapping $q(L)$ on almost all elements of $\mathcal{G}$.
  - If $h, l, h', l'$ are distinct, then $q_h, q_l, q_h'$ are also distinct.
  - $(h, h') \rightarrow (q_h, q_l, q_h')$, $(h, l) \rightarrow (q_h, q_l, q_h)$
  - $(h, l', h') \rightarrow (q_h, q_l, q_h)$, $(l', h') \rightarrow (q_h, q_l, q_h)\delta$.
- The function $q$ maps even slightly different contexts to completely different numbers.
- Using FF, map each unique number $q(L)_k$ to the desired output value $\mathcal{F}(L)_k$.

Alternative Architectures

- Realizing contextual mappings is sufficient to enable universal approximation property.
- Cheaper architectures to implement some forms of contextual mappings?
- Bi-linear projection
  $$B_{\text{Proj}}(X) = X + W_0 \cdot X \cdot W_P$$
  - For random $W_P$, $(X_1 - X_2)W_P$ is dense for sparse $X_1 - X_2 \implies$ a form of "pairwise contextual mapping."
- Depth-wise separable convolutions
  $$S_{\text{FF}}(X) = X + W_0 (X \ast W_C), \quad W_C \in \mathbb{R}^{d \times k}, \quad (X \ast W_C)_{i,j} := X_{i\cdot} \ast (W_C)_{i,j}$$

Performance comparison on BERT BASE

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Avg. Attention</th>
<th>BProj</th>
<th>SepConv</th>
<th>Transformer</th>
</tr>
</thead>
<tbody>
<tr>
<td># params</td>
<td>68.3M</td>
<td>90M</td>
<td>102.5M</td>
<td>110M</td>
</tr>
<tr>
<td>Masked LM acc. (%)</td>
<td>28</td>
<td>59</td>
<td>63</td>
<td>76</td>
</tr>
<tr>
<td>MNLI acc. (%)</td>
<td>66</td>
<td>72.3</td>
<td>73</td>
<td>78.2</td>
</tr>
</tbody>
</table>

Hybrid Transformers

- Modify the first few Transformer blocks.
- Replace self-attention layers with depth-wise separable convolutional layers.