

Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types.  $n$  is the sequence length,  $d$  is the representation dimension,  $k$  is the kernel size of convolutions and  $r$  the size of the neighborhood in restricted self-attention.

| Layer Type                  | Complexity per Layer     | Sequential Operations | Maximum Path Length |
|-----------------------------|--------------------------|-----------------------|---------------------|
| Self-Attention              | $O(n^2 \cdot d)$         | $O(1)$                | $O(1)$              |
| Recurrent                   | $O(n \cdot d^2)$         | $O(n)$                | $O(n)$              |
| Convolutional               | $O(k \cdot n \cdot d^2)$ | $O(1)$                | $O(\log_k(n))$      |
| Self-Attention (restricted) | $O(r \cdot n \cdot d)$   | $O(1)$                | $O(n/r)$            |