Theory of Transformer Expressive Power

Xiyu Zhai

Contents

1 Introduction

In this paper, we study the expressive power of transformer in a constructive manner and at a system level (as opposed to Turing machine level), aiming at giving insight to its strength and weakness and how to build a better system, potentially vastly distinct from deep learning ones.

More specifically, we formulate several typical computational problems of practical importance and try to argue the power of transformers on expressing the solutions of these problems under three criterions:

(i) Accuracy;

(ii) Statistical Efficiency; how many samples needed to learn it in the most optimal case;

(iii) Computational Efficiency;

Remark. *The computation problems of computer vision and natural language parsing (I mean parsing, not including semantics and reasoning) are notoriously hard to formulate accurately in easy ways. However, as people are applying transformers to harder AI problems, many of those can be formulated in a clean way (say theorem proving), or can be seen as a dirty version of a clean problem but not intrinsically different (common sense reasoning).*

It's just hard to enumerate or establish proof for all possible choices of architecture and weights. Instead, we artificially design the architecture and weights in the hope that it will be close to the optimal and we will compare it with the optimal design we could get on tradition CPU/GPU programs. There are two scenarios where the artificial design deviates from the actual way how transformer works,

(i) the transformer in practice doesn't learn as good a set of weights, which means the current transformer has room to improve. But in this way, we give an upper bound

(ii) the transformer in practice learn a better set of weights that we imagined. This will be surprising and highly interesting, and there can be experiments to prove this if this is true by training a small restrive transformer on a domain specific dataset and comparing the performance.

In short, we give upper bounds of the abilities of transformers in various tasks which can be easily invalidated if false.

2 Refined Machine Learning Framework