Journal of Machine Learning **ISSN**: 2790-2048(e), 2790-203X(p)

Why Self-Attention is Natural for Sequence-to-Sequence Problems? A Perspective from Symmetries

Chao Ma * [1] and Lexing Ying † [1]

[1]Department of Mathematics Stanford University, Stanford, CA 94305, USA

Abstract. In this paper, we show that structures similar to self-attention are natural for learning many sequence-to-sequence problems from the perspective of symmetry. Inspired by language processing applications, we study the orthogonal equivariance of seq2seq functions with knowledge, which are functions taking two inputs – an input sequence and a knowledge – and outputting another sequence. The knowledge consists of a set of vectors in the same embedding space as the input sequence, containing the information of the language used to process the input sequence. We show that orthogonal equivariance in the embedding space is natural for seq2seq functions with knowledge, and under such equivariance, the function must take a form close to self-attention. This shows that network structures similar to self-attention are the right structures for representing the target function of many seq2seq problems. The representation can be further refined if a finite information principle is considered, or a permutation equivariance holds for the elements of the input sequence.

Keywords:
Self attention,
Symmetry,
Orthogonal equivariance,
Permutation equivariance.

Article Info.:
Volume: 2
Number: 3
Pages: 194 - 210
Date: September /2023
doi.org/10.4208/jml.221206

Article History:
Received: 06/12/2022
Accepted: 28/08/2023

Communicated by:
Zhi-Qin Xu

1 Introduction

Neural network models using self-attention, such as Transformers [47], have become the new benchmark in the fields such as natural language processing and protein folding. Though, the design of self-attention is largely heuristic, and a theoretical understanding of its success is still lacking. In this paper, we provide a perspective for this problem from the symmetries of sequence-to-sequence (seq2seq) learning problems. By identifying and studying appropriate symmetries for seq2seq problems of practical interest, we demonstrate that structures like self-attention are natural for representing these problems.

Symmetries in learning problems can inspire the invention of simple and efficient neural network structures. This is because symmetries reduce the complexity of the problems, and a network with matching symmetries can learn the problems more efficiently. For instance, convolutional neural networks (CNNs) have seen great success in vision problems, with the translation invariance/equivariance of the problems being one of the main reasons. This is not only observed in practice but also justified theoretically [21]. Many other symmetries have been studied and exploited in the design of neural network

*Corresponding author. chaoma@stanford.edu
†lexing@stanford.edu

models. Examples include permutation equivariance [57] and rotational invariance [8,17], with various applications in learning physical problems. See Section 2.1 for more related works.

In this work, we start by studying the symmetry of seq2seq functions in the embedding space, the space in which each element of the input and output sequences is represented. For a language processing problem, for example, words or tokens are usually vectorized by a one-hot embedding using a vocabulary. In this process, the order of words in the vocabulary should not influence the meaning of input and output sentences. Thus, if a permutation is applied on the dimensions of the embedding space, the input and output sequences should experience the same permutation, without other changes. This implies a permutation equivariance in the embedding space. In our analysis, we consider equivariance under the orthogonal group, which is slightly larger than the permutation group. We show that if a function f is orthogonal equivariant in the embedding space, then its output can be expressed as linear combinations of the elements of the input sequence, with the coefficients only depending on the inner products of these elements. Concretely, let $X \in \mathbb{R}^{d\times n}$ denote an input sequence with length n in the embedding space $\mathbb{R}^d$. If $f(QX)=Qf(X)$ holds for any orthogonal $Q\in\mathbb{R}^{d\times d}$, then there exists a function g such that

$$f(X) = Xg(X^TX).$$

However, the symmetry on the embedding space is actually more complicated than a simple orthogonal equivariance. In Section 3.2, we show that the target function for a simple seq2seq problem is not orthogonal equivariant, because the target function works in a fixed embedding. To accurately catch the symmetry in the embedding space, we propose to study seq2seq functions with knowledge, which are functions with two inputs, $f(X,Z)$, where $X \in \mathbb{R}^{d\times n}$ is the input sequence and $Z \in \mathbb{R}^{d\times k}$ is another input representing our knowledge of the language. The knowledge lies in the same embedding space as X and is used to extract information from X. With this additional input, the symmetry in the embedding space can be formulated as an orthogonal equivariance of $f(X,Z)$, i.e. $f(QX,QZ) = Qf(X,Z)$ for any inputs and orthogonal matrix Q. Intuitively understood, in a language application, as long as the knowledge is always in the same embedding as the input sequence, the meaning of the output sequence will not change with the embedding. Based on the earlier theoretical result for simple orthogonal equivariant functions, if a seq2seq function with knowledge is orthogonal equivariant, then it must have the form

$$f(X,Z) = Xg_1(X^TX, Z^TX, Z^TZ) + Zg_2(X^TX, Z^TX, Z^TZ).$$

If Z is understood as a parameter matrix to be learned, the following subset of this representation:

$$f(X,Z) = Xg(X^TZ)$$

is close to a self-attention used in practice, with Z being the concatenation of query and key parameters. This reveals one possible reason behind the success of self-attention-based models on language problems.

Based on the results from orthogonal equivariance, we further study the permutation equivariance on the elements of the input sequence. Under this symmetry, we show that