UNIVERSIDAD
DE GRANADA

The Attention Mechanism

Francisco Miguel Castro Macías
fcastro@ugr.es

Visual Information Processing Group (VIP)
Department of Computer Science and Artificial Intelligence (DECSAI)
University of Granada (UGR)

May 24, 2023

Overview

1. Introduction
2. Different flavours of Attention: Self-Attention and Cross-Attention
3. Inductive Bias: the need for Positional Encodings

1. Introduction

2. Different flavours of Attention: Self-Attention and Cross-Attention

3. Inductive Bias: the need for Positional Encodings

Motivation

We want to understand what is going on inside a **Transformer**!

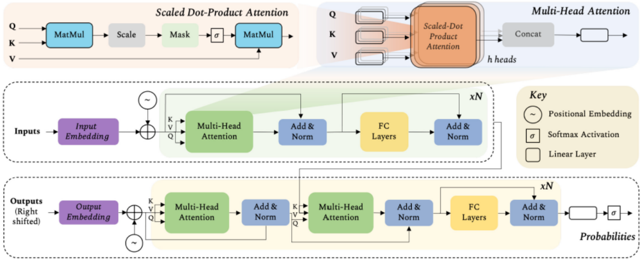

Figure: Architecture of the Transformer Model (Khan et al., 2022).

Useful surveys: Khan et al. (2022); Lin et al. (2022).

1. Introduction

2. Different flavours of Attention: Self-Attention and Cross-Attention

3. Inductive Bias: the need for Positional Encodings

Self-Attention: an initial idea

We are given an initial embedding matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times D}$. There may be some relationship o correlation between the elements of $\mathbf{X}$. We want to compute new representations $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]^\top \in \mathbb{R}^{N \times D}$ exploiting that correlation.

- We compute each $\mathbf{z}_i$ as a convex combination of the elements in $\mathbf{X}$,

$$\mathbf{z}_i = \sum_{j=1}^{N} A(\mathbf{x}_i, \mathbf{x}_j)\mathbf{x}_j,$$

  where $A(\mathbf{x}_i, \mathbf{x}_j) \in (0, 1)$ and $\sum_j A(\mathbf{x}_i, \mathbf{x}_j) = 1$.

- $A(\mathbf{x}_i, \mathbf{x}_j)$ should represent some similarity measure between $\mathbf{x}_i$ and $\mathbf{x}_j$. Let's use a kernel!

$$A(\mathbf{x}_i, \mathbf{x}_j) = \frac{\exp(\mathbf{x}_i^\top \mathbf{x}_j)}{\sum_l \exp(\mathbf{x}_i^\top \mathbf{x}_l)}$$

- ... What is this?

$$\mathbf{Z} = \mathrm{Softmax}(\mathbf{X}\mathbf{X}^\top)\mathbf{X}$$

Self-Attention: adding complexity

$$\mathbf{Z} = \mathrm{Softmax}(\mathbf{X}\mathbf{X}^\top)\mathbf{X} \in \mathbb{R}^{N\times D}$$

Some questions...

- Is the original space rich enough to capture the dependencies / correlations?
- Is the original space rich enough to express the new representations?

Answer: we don't know, so we learn new ones,

$$\mathbf{Z} = \mathrm{Softmax}(\phi_q(\mathbf{X})\phi_k(\mathbf{X})^\top)\phi_v(\mathbf{X}) \in \mathbb{R}^{N\times S},$$
$$\phi_q : \mathbb{R}^{N\times D} \to \mathbb{R}^{N\times L}, \quad \phi_k : \mathbb{R}^{N\times D} \to \mathbb{R}^{N\times L}, \quad \phi_v : \mathbb{R}^{N\times D} \to \mathbb{R}^{N\times S}.$$

Remarks.

- The matrix $A(\mathbf{X}, \mathbf{X}) = \mathrm{Softmax}(\phi_q(\mathbf{X})\phi_k(\mathbf{X})^\top)$ is known as the *attention matrix.*
- Usually, ϕ_q, ϕ_k, ϕ_v are defined element-wise, i.e., $\phi_*(\mathbf{X}) = [\phi_*(\mathbf{x}_1), \ldots, \phi_*(\mathbf{x}_N)]^\top$.

Cross-Attention

The same reasoning can be applied to deal with two differents embbeding matrices, $\mathbf{X} \in \mathbb{R}^{N \times D_{\mathbf{X}}}$ and $\mathbf{Y} \in \mathbb{R}^{M \times D_{\mathbf{Y}}}$. This is called *Cross-Attention*,

$$\mathbf{Z} = \text{Softmax}(\phi_q(\mathbf{X})\phi_k(\mathbf{Y})^\top)\phi_v(\mathbf{Y}) \in \mathbb{R}^{N \times S},$$

$$\phi_q \colon \mathbb{R}^{N \times D_{\mathbf{X}}} \to \mathbb{R}^{N \times L}, \quad \phi_k \colon \mathbb{R}^{M \times D_{\mathbf{Y}}} \to \mathbb{R}^{N \times L}, \quad \phi_v \colon \mathbb{R}^{M \times D_{\mathbf{Y}}} \to \mathbb{R}^{M \times S}.$$

Remark. Let

$$\mathbf{X} = 1 \in \mathbb{R}^{1 \times 1}, \mathbf{Y} \in \mathbb{R}^{M \times D},$$
$$\phi_q = \text{Id}, \quad \phi_v = \text{Id},$$
$$\phi_k(\mathbf{Y}) = \tanh\left(\mathbf{Y}\mathbf{V}\right)\mathbf{w}, \quad \mathbf{w} \in \mathbb{R}^L, \mathbf{V} \in \mathbb{R}^{D \times L}$$

Then, we recover the Attention-based pooling used in Ilse et al. (2018).

Scaled Dot-Product Attention

Self-Attention and Cross-Attention are just realizations of the *Scaled Dot-Product Attention* mechanism proposed in Vaswani et al. (2017).

Scaled Dot-Product Attention, Vaswani et al. (2017).

Let $\mathbf{Q} \in \mathbb{R}^{N \times d}, \mathbf{K} \in \mathbb{R}^{M \times d}$ and $\mathbf{V} \in \mathbb{R}^{M \times S}$. The output of the Scaled Dot-Product Attention mechanism is

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(d^{-1/2}\mathbf{Q}\mathbf{K}^\top\right)\mathbf{V}.$$

The output of h-head Attention mechanism is

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right)\mathbf{W}^O,$$

$$\text{head}_i = \text{Attention}\left(\mathbf{Q}\mathbf{W}_i^{\mathbf{Q}}, \mathbf{K}\mathbf{W}_i^{\mathbf{K}}, \mathbf{V}\mathbf{W}_i^{\mathbf{V}}\right)$$

1. Introduction

2. Different flavours of Attention: Self-Attention and Cross-Attention

3. Inductive Bias: the need for Positional Encodings

Inductive bias

Question. What kind of functions do Self-Attention layers represent?

Suppose that ϕ_q, ϕ_k, ϕ_v are defined element-wise, i.e., $\phi_*(\mathbf{X}) = [\phi_*(\mathbf{x}_1), \ldots, \phi_*(\mathbf{x}_N)]^\top$.

$$\mathbf{Z} = \mathrm{SA}(\mathbf{X}) = \mathrm{Softmax}(\phi_q(\mathbf{X})\phi_k(\mathbf{X})^\top)\phi_v(\mathbf{X}),$$

$$\mathbf{z}_i = \mathrm{SA}(\mathbf{x}_i) = \sum_{j=1}^{N} A(\phi_q(\mathbf{x}_i), \phi_k(\mathbf{x}_j))\phi_v(\mathbf{x}_j).$$

Then, under a permutation of the elements in $\mathbf{X}$,

- $\mathbf{x} \mapsto \mathrm{SA}(\mathbf{x})$ is *invariant*.
- $\mathbf{X} \mapsto \mathrm{SA}(\mathbf{X})$ is *equivariant*.

Therefore, Self-Attention layers are useful for learning representations of *unordered sets* (Bronstein et al., 2021). Then, why do Transformers excel when dealing with sequences?

Positional Encoding

Positional information is injected into the embeddings using a positional encoding transformation,

$$\widehat{\mathbf{x}}_i = \mathrm{PE}(i, \mathbf{x}_i, \mathbf{X}) \in \mathbb{R}^{D'}.$$

The input embedding matrix $\mathbf{X}$ is transformed into $\widehat{\mathbf{X}} = [\widehat{\mathbf{x}}_1, , \ldots, \widehat{\mathbf{x}}_N]^\top$. Then, under a permutation of the elements in $\mathbf{X}$, $\mathbf{x} \mapsto \mathrm{SA}(\widehat{\mathbf{x}})$ is no longer *invariant* and $\mathbf{X} \mapsto \mathrm{SA}(\widehat{\mathbf{X}})$ is no longer *equivariant*.

Thank you!

References I

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI Open*, 2022.

References II

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.