Assignment 6: Transformers

CSE 256: Statistical NLP: Spring 2022

University of California, San Diego

Released: May 23, 2022

Due: May 30, 2022

In this assignment, you will answer conceptual questions about Attention, and Transformers. Unless stated, all questions are asking about content covered in class. For most questions, we expect your answer to be about 2-3 sentences, but can be longer if desired.

Part 1 Encoder-Decoder (4 Points)

Translation is a sequence-to-sequence (seq-to-seq) problem: both the input and output are sequences (of words). A common neural architecture used for seq-to-seq problems is the encoder-decoder model, depicted in Figure1. In its original formulation, the encoder-decoder model consists of two recurrent neural networks (RNNs). The encoder RNN compresses the input sequence into a fixed-length vector, represented by the final encoder hidden state. The decoder RNN conditions on this vector to produce the translation.

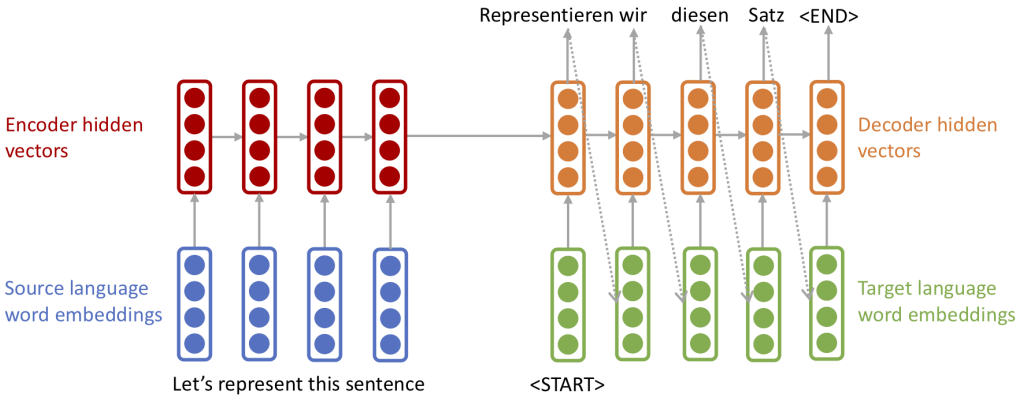

Figure 1: Encoder-Decoder architecture

Q1 How do you think the architecture in Figure1 will perform on long sequences, and why? Consider the amount of information the decoder gets to see about the input sequence (*2 points*).

Q2 What modification can we make to this architecture to improve its performance on long sequences? Explain how this modification facilitates more accurate translation on long sequences. (*2 points*)

Part 2 Transformers vs RNNs (2 Points)

Q3 Today, Transformers have replaced RNNs in the encoder-decoder architectures. Transformers get rid of the recurrence that is central to RNNs. Instead, they rely completely on attention. In self-attention, each element in the sentence attends to other elements, resulting in context-sensitive representations.

Transformers have allowed the NLP community to train on larger datasets than was once possible, what aspect of the Transformer makes this possible in contrast to RNNs? (*2 Points*)

Part 3 Self-Attention and Variants (6.5 Points)

Recall that attention can be viewed as an operation on a query $q \in \mathbb{R}^d$, a set of value vectors $\{v_1, \ldots, v_n\}, v_i \in \mathbb{R}^d$, and a set of key vectors $\{k_1, \ldots, k_n\}, k_i \in \mathbb{R}^d$, specified as follows:

$$c = \sum_{i=1}^{n} v_i \alpha_i$$

$$\alpha_i = \frac{\exp\left(k_i^\top q\right)}{\sum_{j=1}^{n} \exp\left(k_j^\top q\right)}$$

with α_i termed the "attention weights". The output $c \in \mathbb{R}^d$ is an average over the value vectors weighted with respect to α_i.
Here the attention function that computes α_i, is the **dot product**, applied to the query and the key vector $\left(k_i^\top q\right)$.
Self-attention in the original Transformer has quadratic computational complexity with respect to the sequence length. This is because, for a sentence of length n, the dot product between n^2 pairs of word vectors is computed. The full self-attention matrix with every column filled, $O\left(n^2\right)$, is depicted in Figure 2(a).
To reduce runtime, variants have been proposed to sparsify the full self-attention matrix according to an "attention pattern" specifying pairs of input locations attending to one another. For example, the *sliding window attention* uses a fixed-size window surrounding each token, see Figure 2(b) (reference Beltazy et al, 2020).

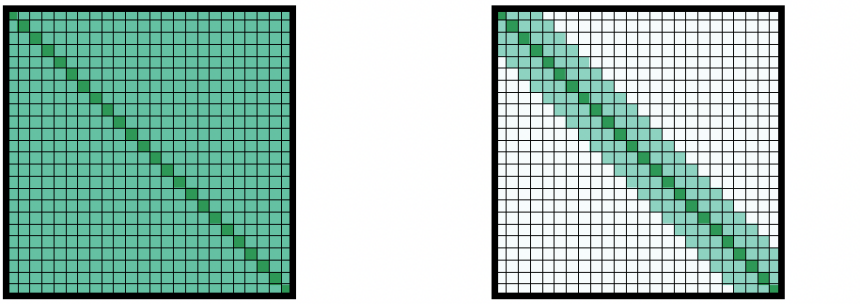

Figure 2: The full self-attention pattern (a) and a window attention pattern(b).

Q4 Under what conditions might the performance of window attention be at a strong disadvantage in comparison to full self-attention? You can reason using specific examples, or you can provide your argument in more abstract terms. (*2 Points*)

Q5 Discuss another self-attention pattern (other than full-self attention and window attention). This can be your own novel pattern, we encourage you to think of your own, but you can also consult the literature for help. Explain one strength and one potential limitation of your chosen attention pattern. You may choose to show a visual similar to Figure 2 to illustrate your attention pattern (*4.5 Points*)

Part 4 Attention Functions (Bonus: 1.5 Points)

In Part 3, we used dot product attention. Name an alternative attention function, and explain its strengths and weaknesses with respect to dot product attention. (**1.5 Points**)

Submission Instructions

Submit your work on Gradescope.

- **Code:** There is no code for this homework
- **Report:** Submit your report, it should be **1-2 pages long, in pdf** (reasonable font sizes).

References

I. Beltagy, M. Peters, A. Cohan. Longformer: The Long-Document Transformer2020.