

EECS 182 Deep Neural Networks

Spring 2023 Anant Sahai

Discussion 10

1. Transformers and Pretraining

Transformer Architecture is illustrated in the schematic below.

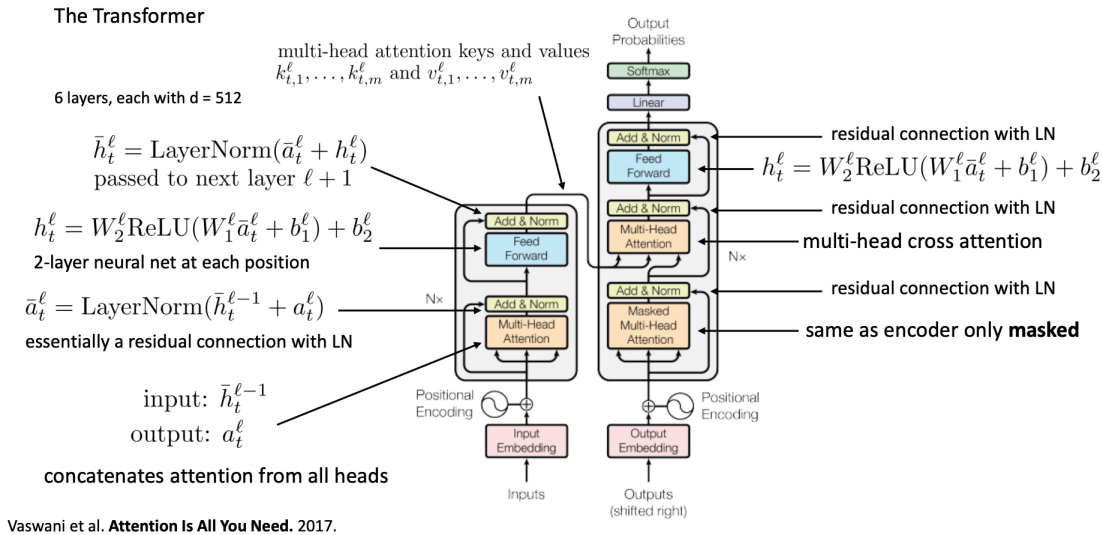


Figure 1: Overview of Transformer architecture

- (a) **Why do we need positional encoding? Describe a situation where word order information is necessary for the task performed.**
- (b) When using an absolute positional encoding (e.g. sinusoids at different frequencies like hands of a clock), we can either add it to the input embedding or concatenate it. That is, if x_i is our word embedding and p_i is our position embedding, we can either use $z = x_i + p_i$ or $z = [x_i, p_i]$. Consider a simple example where the query and key for the attention layer are both simply $q = k = z$. **If we compute a dot-product of a query with another key in the attention layer, what would be the result in either case? Discuss the implications of this.**

- (c) It turns out we can extend the self-attention mechanics to have relative positions matter without cross terms and without having to explicitly concatenate (and thereby increase the length of) two kinds of embeddings.
- Relative position embedding explicitly adds a learnable set of biases π_{i-j} to the dot-product scores before the softmax operation. **For what $\pi_{i,j}$ would we get the same behavior from attention as concatenating the position embeddings $q_i^{(\text{pos})}, k_j^{(\text{pos})}$ to both the query q_i and the keys k_j ?**
 - If T is the maximum context-length and we have m attention heads, **how many extra parameters have to be learned if we insist that these learned (per-head) attention biases $\pi_{i,j}$ must only depend on relative position?**
- (d) **What is the advantage of multi-headed attention?** Give some examples of structures that can be found using multi-headed attention.
- (e) Let's say we're using argmax attention, which uses argmax rather than softmax, like we saw on the midterm. **What is the size of the receptive field of a node at level n ...**
If we have only a single head?
If we have two heads?
If we have k heads?
- (f) For input sequences of length M and output sequences of length N , **what are the complexities of (1) Encoder Self-Attention (2) Cross Attention (3) Decoder Self-Attention.** Let k be the hidden dimension of the network.
- (g) **True or False: With transformer masked autoencoders, masking out a token typically involves replacing both the token value and the positional encoding at an index with a special "mask" token.**

(h) A group of CS 182 students are creating a language model, and one student suggests that they use random text from novels for pre-training. Another student says that this is just arbitrary text isn't useful because there aren't any labels. **Who's right and why?**

(i) **Would an encoder model or a encoder-decoder model be better suited for the following tasks?**

Summarizing text in an article

Classify written restaurant reviews by their sentiment

Identifying useful pages when retrieving web search results

Translating one language to another

Contributors:

- Kevin Li.
- Anant Sahai.
- Olivia Watkins.
- Jerome Quenum.
- Saagar Sanghavi.
- CS 182/282A Staff from previous semesters.