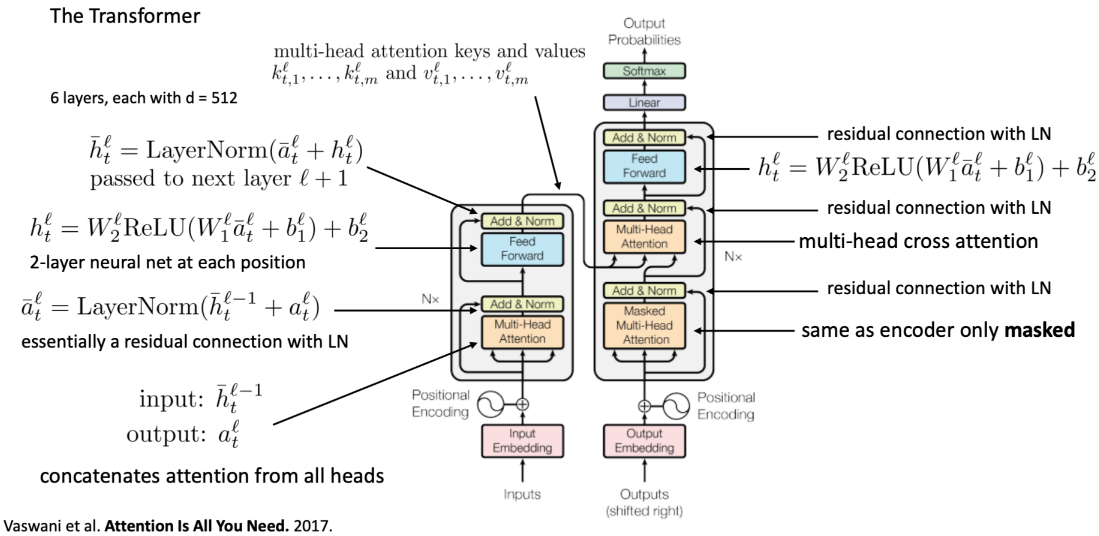
EECS 182 Deep Neural Networks
Fall 2022 Anant Sahai

Discussion 9

1. Tranformers and Pretraining

Tranformer Architecture is illustrated in the schematic below.


The Transformer
6 layers, each with d = 512
multi-head attention keys and values
$k_{t,1}^{\ell}, \ldots, k_{t,m}^{\ell}$ and $v_{t,1}^{\ell}, \ldots, v_{t,m}^{\ell}$
$\bar{h}_t^{\ell} = \text{LayerNorm}(\bar{a}_t^{\ell} + h_t^{\ell})$
passed to next layer $\ell + 1$
$h_t^{\ell} = W_2^{\ell}\text{ReLU}(W_1^{\ell}\bar{a}_t^{\ell} + b_1^{\ell}) + b_2^{\ell}$
2-layer neural net at each position
$\bar{a}_t^{\ell} = \text{LayerNorm}(\bar{h}_t^{\ell-1} + a_t^{\ell})$
essentially a residual connection with LN
input: $\bar{h}_t^{\ell-1}$
output: a_t^{ℓ}
concatenates attention from all heads
Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
N×
N×
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)
residual connection with LN
$h_t^{\ell} = W_2^{\ell}\text{ReLU}(W_1^{\ell}\bar{a}_t^{\ell} + b_1^{\ell}) + b_2^{\ell}$
residual connection with LN
multi-head cross attention
residual connection with LN
same as encoder only **masked**

Vaswani et al. **Attention Is All You Need.** 2017.

Figure 1: Overview of Transformer architecture

(a) Why do we need positional encoding? Describe a situation where positional encoding is necessary for the task performed.

(b) When using the positional encoding, 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 input $z = [x_i, p_i]$ 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) What is the advantage of multi-headed attention? Give some examples of structures that can be found using multi-headed attention.

(d) 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?

(e) 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.

(f) 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.

(g) 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?

(h) Would an encoder model or a seq-to-seq 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

(i) What are the pros and cons of each of the discussed pretrained language models: ELMo, BERT, and GPT? In which situations is each type of model most useful for?