Q&A

Dashygo

September 4, 2024

Q1 为什么bert不能像gpt一样做文本生成？

A1 The original model of BERT is pre-trained for MLM and NSP tasks,which means that BERT is better at NLU(*Natural Language Understanding*).What's more, BERT serves as a *bi-directional language model* ,which is somewhat contradictory to generative tasks(Because in generative tasks,we can only see the words that comes before we make predictions),while you can also take the last word for [MASK] for BERT to predict ,then it may not be comparable to models that are specially designed for this purpose).

Different from BERT, GPT is targeted at *auto-regression* optimization task below:

$$max\, L_1(\mathcal{U}) = \sum_i log\, P(u_i|u_{i-k}, ..., u_{i-1}; \theta)$$

in which k is the size of context window.

As we can see,it is literally for generative tasks just in mathematics.(maximum likelihood for the prediction of next token)

Moreover ,GPT has multi-headed-attention mechanism with *mask*,which the prevent follow-up tokens from influencing predictions,owing to the decoderblock from original transformer.

Q2 对于decoder-only的模型，它是由tokenizer,embedding层, transformer block,lm_head,请你简单说明一下这4个部分分别在做什么？token是一个什么东西？

A2 *Token* is a fundamental unit in text.It can be words,letters,punctuation marks,or some special pieces such as "[MASK]"(in BERT for example),"[PAD]"(padding for sentences with different lengths), "##ing"(WordPiece) etc.

Tokenizer:Vectorize the text into vectors that computer can deal with or just split text into tokens.

Embedding:Turn vectorized text into high-dimentional embed_vectors,in order to distinguish the characteristics of tokens,just like $E(man)-E(woman) \approx E(king)-E(queen)$ as someone puts it.There can be some different embedding layers like positional embedding layers in transformer.Or in some cases,we can also find the similarities between two words with Embedding layer.Embedding is often learnable and serves under some learning strategies.

Transformer_block:Classical DecoderBlock in transformer for n times with *Masked-Multi-Headed-Attention*,*LayerNorm* with *Residual*，*FFN(Feed Forward Networks)*.

Lm_head:A linear layer, mapping the outputs of decoderblock to vocab_dim (same size of tokens),then sofmax or any other kinds of normalization methods to get output probabilities.

Q3 为什么decoder-only的模型的数量远远多于Encoder-Decoder模型？明明二者都可以做文本生成

A3 This is somewhat hard to conclude.I suppose in most circumstances, with same training cost and numbers of parameters,it will be more effecient to build a Decoder-Only models than Encoder-Decoder models, while this may hard to explain in mathematics, but may be proved by engineering practices.

Aside from this, there is an explanition in mathematics which makes some sense: that is the loss of the rank of pure bi-directional attention.While in single directional attention there won't be this problem because after softmax, weight matrix is a *upper triangular matrix* which has a rank same as the dimension of matrix.But in bi-directional attention this is not always the case.