CLIP text-embedding 问题

hf transformers clip 的 text embedding 实现：

(代码链接 https://github.com/huggingface/transformers/blob/af3de8d87c717c4bb090f037d0d89413c195a42f/src/transformers/models/clip/modeling_clip.py#L724)

在 textEncoder 模块

首先输出的 last_hidden_state 形状是 [batch, query_length, d_model]

最终输出的 pooled_output 形状是 [batch, d_model]

其中 pooled_output 只是选取了 last_hidden_state 的 eos_token 对应的 embedding, 其余的抛弃

问题：

用 eos_token embedding 代表整个句子 embedding 的原因是什么？ eos 不应该只代表句子结束吗（即使有 attention 相互牵连，也不应代表句子语义吧）？

其他资料：

openai/CLIP 官方代码表述：take features from the eot embedding (eot_token is the highest number in each sequence)

(代码链接 https://github.com/openai/CLIP/blob/a1d071733d7111c9c014f024669f959182114e33/clip/model.py#L353)

CLIP 原论文表述：

The text sequence is bracketed with [SOS] and [EOS] tokens and the activations of the highest layer of the transformer at the [EOS]

token are treated as the feature representation of the text

(论文链接 https://arxiv.org/pdf/2103.00020.pdf)

解答：

1、BERT 里面此 token 叫做 [CLS]， 代表句子“整体信息”分类。

2、操作叫 pooler。

3、attention 操作本身就是权重后相加。

4、也有其他 pool方法 pooling_mode_cls_token | pooling_mode_mean_tokens | pooling_mode_max_tokens | pooling_mode_mean_sqrt_len_tokens 1_Pooling/config.json
https://huggingface.co/consciousAI/cai-lunaris-text-embeddings/blob/main/1_Pooling/config.json

BERT https://arxiv.org/pdf/1810.04805.pdf

“The first token of every sequence is always **a special classification token ([CLS])**. The final hidden state

corresponding to this token **is used as the aggregate sequence representation** for classification tasks.”