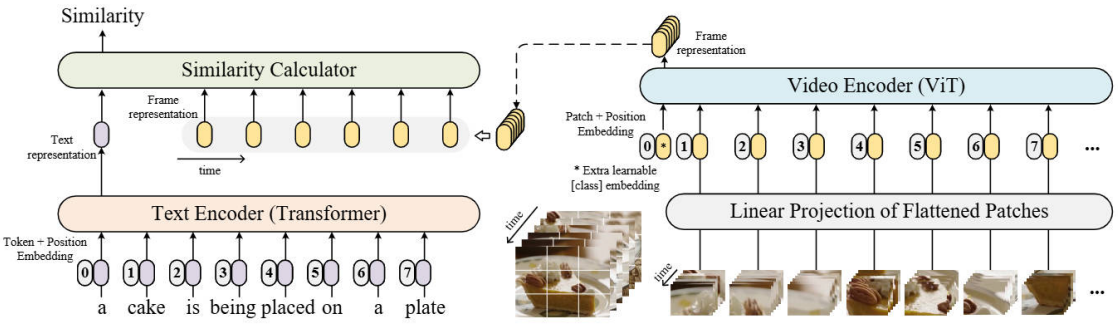
CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval

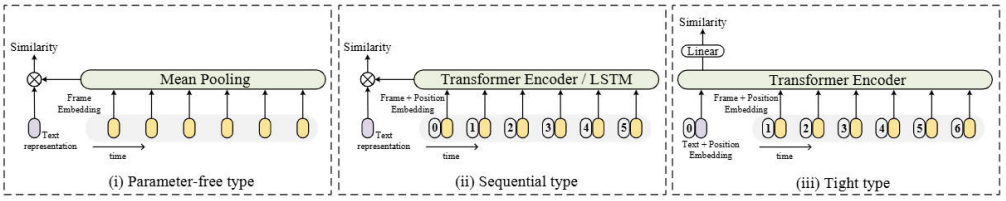
动机：

将CLIP迁移到视频处理领域，但在视频领域中，还需要考虑到temporal dependency

Architecture：



(a) Main structure



(b) Similarity calculator

Video Encoder：ViT

$$ViT(V_i = \{V_i^1, V_i^2, \ldots, V_i^{|V_i|}\})$$

[class] token做为video representation

它的Linear Projection of Flattened Patches有两种类型，2D linear和3D linear，区别在于前者的convolution核为[hxw]，后者的convolution核为[txhxw]，考虑了时序

Text Encoder：Transformer

$$t_j \in \tau$$

[EOS] token做为text representation

Similarity Calculator：三种

这个很重要，因为CLIP是image-text pairs，通过这个similarity calculator可以转移到video-text

1. Parameter-free：Mean Pooling，虽然有丢失时序的缺陷，但是仍然被广泛使用，后两个要在下游任务数据集较大时，才效果比较好

2. Sequential type：考虑sequential information，有两种方法，LSTM或Transformer

3. Tight type：将两种模态深度交叉

Insight：

1. image-text pairs数据能用来提升视频领域中的模型

2. 要post-pretrain，即从“image-text pairs”到“video-text pairs”

3. CLIP用在视频领域中，对learning-rate十分敏感