- 架构一览
  - CLIP模型（顶层）
    - 核心功能：学习图像与文本的联合表示，计算跨模态相似度
    - 关键组件：视觉编码器、文本编码器、跨模态交互模块
    - 输出：图像-文本相似度分数（logits_per_image、logits_per_text）
  - 视觉编码器（visual）
    - ModifiedResNet（残差网络变体）
      - Stem模块（输入处理）
        - conv1：3x3卷积（输入3通道→width//2通道，步长2，填充1）
        - bn1：BatchNorm2d（归一化）
        - relu1：ReLU（激活）
        - conv2：3x3卷积（width//2→width//2通道，填充1）
        - bn2：BatchNorm2d
        - relu2：ReLU
        - conv3：3x3卷积（width//2→width通道，填充1）
        - bn3：BatchNorm2d
        - relu3：ReLU
        - avgpool：AvgPool2d（步长2，压缩特征图）
      - 残差层（layer1~layer4）
        - 每层由多个`Bottleneck`块组成（数量由`layers`参数指定）
        - 每个`Bottleneck`块
          - conv1：1x1卷积（降维，inplanes→planes）
          - bn1：BatchNorm2d
          - relu1：ReLU
          - conv2：3x3卷积（特征提取，planes→planes，填充1）
          - bn2：BatchNorm2d
          - relu2：ReLU
          - avgpool：AvgPool2d（步长>1时使用，否则为Identity）
          - conv3：1x1卷积（升维，planes→planes×4，expansion=4）
          - bn3：BatchNorm2d
          - downsample（可选）：当步长>1或通道不匹配时
            - AvgPool2d（下采样）
            - 1x1卷积（匹配通道）
            - BatchNorm2d
          - relu3：ReLU（残差连接后激活，out += identity）
      - AttentionPool2d（注意力池化）
        - positional_embedding：位置嵌入参数
        - k_proj：Linear（键投影）
        - q_proj：Linear（查询投影）
        - v_proj：Linear（值投影）
        - c_proj：Linear（输出投影）
        - 多头自注意力计算：以全局平均特征为查询，结合位置嵌入
    - VisionTransformer（视觉Transformer）
      - conv1：patch卷积（3通道→width通道，kernel=patch_size，步长=patch_size，分割图像为patch）
      - class_embedding：类嵌入参数（类似[CLS]标记）
      - positional_embedding：patch位置嵌入参数
      - ln_pre：LayerNorm（Transformer前归一化）
      - transformer（Transformer编码器）
        - ResidualAttentionBlock
          - 由多个`ResidualAttentionBlock`组成（数量由`layers`参数指定）
          - 每个`ResidualAttentionBlock`
            - ln_1：LayerNorm（注意力前归一化）
            - attn：MultiheadAttention（多头自注意力）
            - ln_2：LayerNorm（前馈网络前归一化）
            - mlp：前馈网络
              - c_fc：Linear（升维，d_model→d_model×4）
              - gelu：QuickGELU（激活）
              - c_proj：Linear（降维，d_model×4→d_model）
        - ln_post：LayerNorm（Transformer后归一化）
        - proj：Linear（投影至输出维度，width→output_dim）
  - 文本编码器
    - transformer（Transformer编码器）
      - transformer
        - 结构同VisionTransformer的transformer，由多个`ResidualAttentionBlock`组成
        - 注意力掩码：因果掩码（上三角矩阵，确保自回归性）
      - token_embedding：Embedding（词嵌入，vocab_size→transformer_width）
      - positional_embedding：位置嵌入参数（长度=context_length）
      - ln_final：LayerNorm（Transformer后归一化）
      - text_projection：Linear（文本特征投影至公共嵌入空间，transformer_width→embed_dim）
  - 跨模态交互
    - 特征归一化：图像/文本特征均做L2归一化
    - 温度系数：logit_scale（参数，控制相似度缩放，exp后使用）
    - 相似度计算：图像特征 × 文本特征转置 × 温度系数，得到双向logits