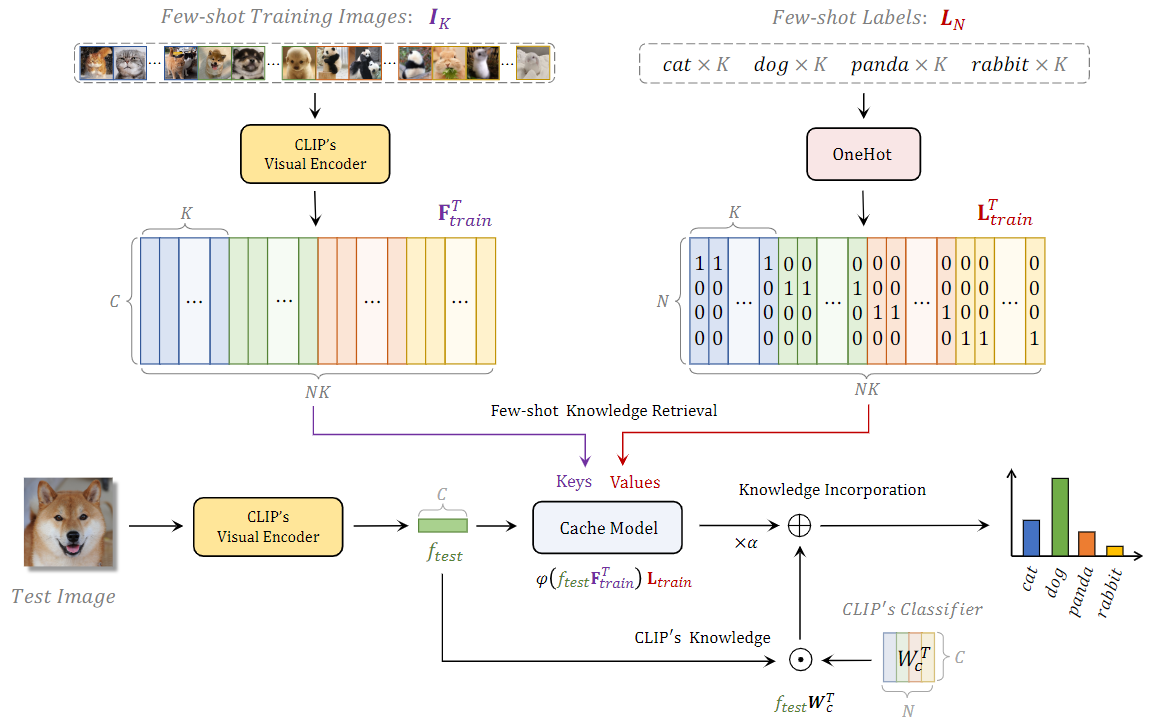
**Tip-Adapter: Training-free Adaption of CLIP for Few-shot Classification**

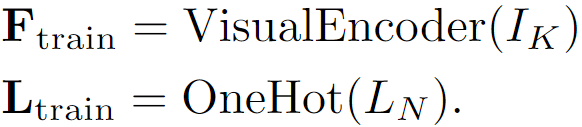
<https://arxiv.org/abs/2207.09519> Eccv2022

为了提高few-shot分类效果，提出了free-training的tip-adapter及其微调变体。Tip-adapter以非参数的方式从few-shot训练集中构造键值缓存模型。而如果可以训练，Tip-Adapter-F通过快速fine-tuning微调缓存键，进一步改善了效果。



**Tip-adapter**

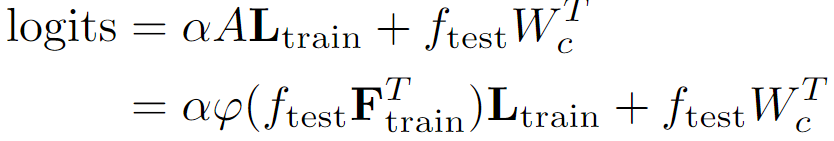
对N类每类K张图，利用CLIP的预训练的视觉编码器提取其c维L2归一化特征，并将标签转换为N维onehot向量。对于所有NK训练样本，Ftrain∈RNK×C和Ltrain∈RN K×N，前者作为键值缓存模型中的keys，后者作为values。



建立缓存模型后，在inference阶段，首先由CLIP的resnet50视觉编码器提取图像的L2归一化特征ftest∈R1×C，并作为query从cache model中检索。Query和key之间的关系：ftestFTtrain等价于余弦相似度，采用指数函数将相似度转换为非负值，β调节相似度的锐度。



A∈R1×N K，attn matrix A去加权value得到ALtrain∈R1×N。除了从cache model中检索到的few shot信息外，预训练的CLIP先验知识由ftestW T c∈R1×N计算，其中Wc为预训练的文本编码器transformer生成的CLIP分类器的权值。Tip-Adapter测试图像的输出为

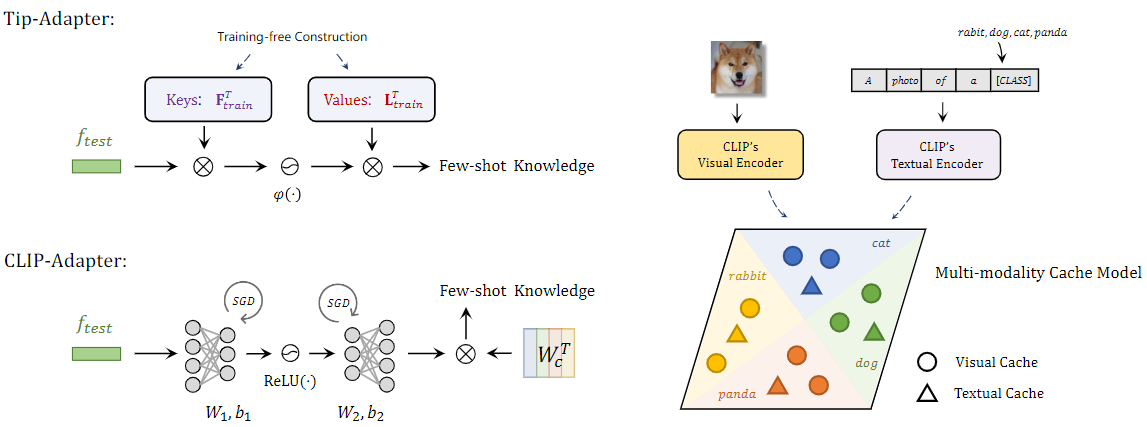


其中：

**Tip-adapter-F**

在shots较多时，未经训练的Tip-Adapter逐渐比不上训练了的CoOp和CLIP-Adapter。为了在保持效率的同时提高指标，提出Tip-Adapter-F，将cache model中的key作为可学习参数的初始化，并通过SGD微调。就是让缓存的键Ftrain可训练，固定了Ltrain和预训练的CLIP的两个编码器。更新缓存模型中的键值可以让相似度矩阵A更准确，从而inference阶段检索更准。而为了更好地记忆类别信息，缓存模型中的值，需要保持固定。Tip-Adapter-F仅在ImageNet上进行20个epoch的微调指标最好，而CoOp和CLIP-Adapter要训练200个epoch。

**与clip adapter区别**



左边两张图显示了tip-adapter的检索cache model信息分支和clip-adapter加了可学习的adapter区别。右图是与以前仅使用视觉特征缓存的网络不同，Tip-Adapter通过CLIP的编码器同时缓存视觉和文本知识。

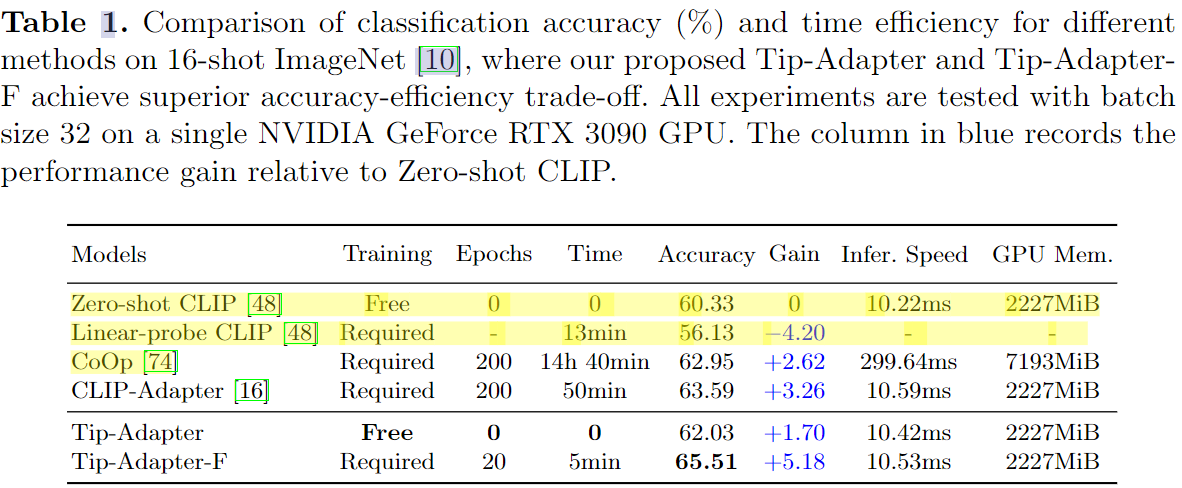
CLIP- adapter将一个双层MLP加到预训练的固定权重的CLIP模型上，并通过SGD优化其参数。对于一个inference阶段的图像，首先由CLIP预先训练的视觉编码器获得feature。然后，接上带有随机初始化参数W1, b1, W2, b2的基于mlp的adapter，输出特征：

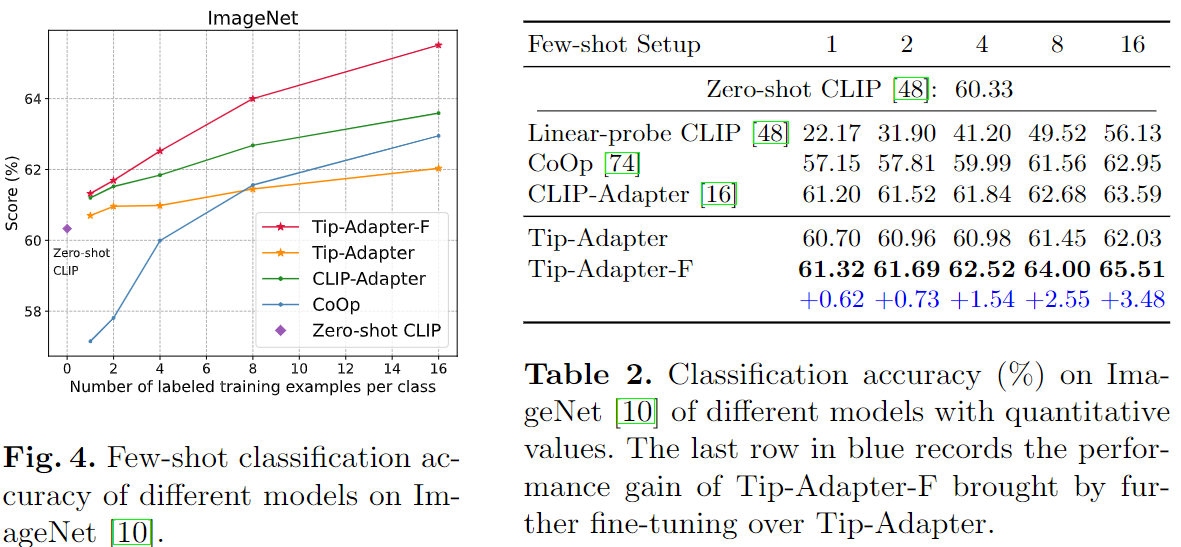
 

其中为MLP中的激活函数，α∈[0,1]。其中W T c为CLIP分类器的权值，会将其与CLIP的W T c相乘，得到最终分类分数。

**超参数设置**：Cache size:16，β：5.5，α：1.0

**结果：**





**问题：**

1. Tip-adapter-F效果比Tip-adapter效果好的原因？

只让cache model的key可训练，其他固定，这样可以更新缓存模型中的key，让attn matrix更准确，并且有clip的预训练编码器输出的feature来作为cache model的key初始化，有良好的初始参数，可以加快收敛。