(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117036901 A
(43)申请公布日 2023.11.10

(21)申请号 202310867841.1

(22)申请日 2023.07.16

(71)申请人 西北工业大学
地址 710072 陕西省西安市友谊西路127号

(72)发明人 王鹏 付铭禹 李煜堃 索伟 张艳宁

(74)专利代理机构 西安凯多思知识产权代理事务所(普通合伙) 61290
专利代理师 刘涛

(51)Int.Cl.
G06V 10/82(2022.01)
G06V 10/764(2022.01)
G06V 10/774(2022.01)
G06N 3/045(2023.01)
G06N 3/0895(2023.01)
G06N 3/09(2023.01)

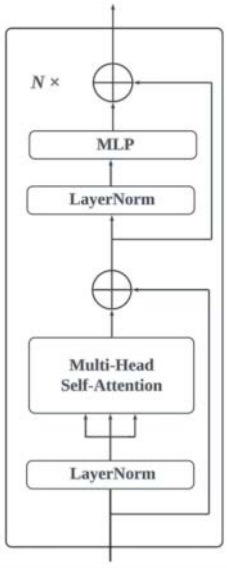
权利要求书1页 说明书4页 附图2页

(54)发明名称

一种基于视觉自注意力模型的小样本微调方法

(57)摘要

本发明公开了一种基于视觉自注意力模型的小样本微调方法,采用在大规模数据集上进行预训练和在小样本任务上进行微调的流程,视觉自注意力模型被用作主干网络,同时构建一个可学习的转换模块norm adapter,由两个向量组成,用于校正原始视觉自注意力模型归一化层的增益和偏置,norm adapter位于视觉自注意力模型ViT的所有归一化层之后,通过逐元素的乘法和加法实现;在预训练期间,使用大规模数据集上以全监督或自监督方式训练的主干网络;在微调过程中,采用原型网络ProtoNet分类头。本发明计算简便,通过逐元素相乘和相加即可实现,因此占用的存储和计算资源比较少,有利于预训练模型投入实际应用场景。



CN 117036901 A

1.一种基于视觉自注意力模型的小样本微调方法,其特征在于,包括如下步骤:

步骤1:构建主干网络;

采用改进的视觉自注意力模型ViT作为主干网络;

原始视觉自注意力模型由一个补丁嵌入层和N个transformer层组成;经过补丁嵌入层,输入图像被编码成一定数量的token向量,再与位置编码相加后,输入的token向量连同CLS token被送入N个transformer层中;最终,经过N个transformer层和一个归一化层LayerNorm后,CLS token用于分类或其他目的;每个transformer层包含两个归一化层LayerNorm,一个MLP块和一个多头自注意力块MHSA;

构建一个可学习的转换模块,由两个向量组成,用于校正原始视觉自注意力模型归一化层LayerNorm的增益gain和偏置bias,将可学习的转换模块称为norm adapter;所述norm adapter位于视觉自注意力模型ViT的所有归一化层之后,通过逐元素的乘法和加法实现;如公式(1)所示,Scale和Shift分别是norm adapter的两个可学习向量,y是归一化层的输出,⊙代表逐元素乘法:

$$h = Scale \odot y + Shift \qquad (1)$$

norm adapter的参数Scale和Shift的结构与归一化层的参数增益gain和偏置bias相同,分别被初始化为全1和全0的向量;在微调时,只有参数Scale和Shift被更新,其他参数在预训练后被冻结,不进行优化;

步骤2:在预训练期间,使用大规模数据集上以全监督或自监督方式训练的主干网络;

步骤3:在微调过程中,采用原型网络ProtoNet分类头;该分类头根据查询图像和原型在嵌入空间中的距离,产生一个概率分布,如公式(2)所示:

$$p_{\phi}(y=k \mid x)=\frac{\exp(-d(f_{\phi}(x),c_{k}))}{\sum_{k'}\exp(-d(f_{\phi}(x),c_{k'}))} \qquad (2)$$

其中,f_{ϕ}是主干网络,将输入编码到特征空间;c_k为类别k的原型,是属于类别k的特征的平均值;d是度量函数;具体来说,各个类别的原型是由支持集中每类样本求均值计算出来的,并将进行数据增强后的支持集作为伪查询集,然后由原型和伪查询集之间的余弦距离计算损失,更新参数;

损失函数选择交叉熵损失。

2.根据权利要求1所述的一种基于视觉自注意力模型的小样本微调方法,其特征在于,所述自监督方式,采用DINO和MOCO v3算法在ImageNet-1K数据集上训练主干网络;所述全监督方式,主干网络是在ImageNet-21K数据集上训练得到的。

3.根据权利要求1所述的一种基于视觉自注意力模型的小样本微调方法,其特征在于,所述度量函数使用余弦距离。

一种基于视觉自注意力模型的小样本微调方法

技术领域

[0001] 本发明属于模式识别技术领域,具体涉及一种基于视觉自注意力模型的小样本微调方法。

背景技术

[0002] 预训练模型被广泛用于自然语言处理(NLP)和计算机视觉(CV)领域,并大大改善了下游任务的性能。因此,预训练-微调范式已经被大家普遍接受,特别是在视觉注意力模型(ViT)兴起之后。由于预训练模型的规模很大,如何在有限的计算和存储开销下将预训练知识有效地迁移到下游任务中,仍在研究之中。已经有一些方法被提出来解决这个问题,称为参数高效微调(PEFT)方法,如:Adapter,bias-tuning,vis ual prompt tuning等等。

[0003] 然而,关于参数高效微调方法在小样本图像分类中的研究却很少。小样本图像分类是小样本学习(few-shot learning)的一项基本任务。小样本学习可以通过模仿人类智能,凭借少量样本泛化到全新的概念,以此来扩大深度学习模型的应用范围。在小样本的设置中,测试数据会被划分成许多任务,每个任务由两部分组成:支持集和查询集,支持集包含N*K个带标注样本,即N个类别的数据,每类有K个样本,这样的小样本任务被称为“N-way K-shot”形式的;查询集包含N*Q数量的样本,用于评估模型。

[0004] 最近,Shell等人首次将预训练模型引入小样本分类领域。他们采用了预训练、元训练,最后微调的流程。首先模型在大规模数据集上(如ImageNet数据集)进行预训练,然后在目标域的基类数据上进行元训练,最后在微调过程中,使用少量样本对模型的所有参数进行更新(full-tuning)。预训练-元训练-微调的流程极大地提高了模型的性能。然而,用于元训练的目标域的基类数据并不容易获得,大多数情况下,只有极少量的带标注样本可以获取到。因此,在这种情况下无法进行元训练,而凭借少量样本对模型的所有参数进行更新(full-tuning)不能充分利用预训练知识。而且,更新全部参数带来的计算和存储开销很大,严重限制了其应用场景。因此,如何在小样本情况下进行高效微调仍然是一个开放的问题。

发明内容

[0005] 为了克服现有技术的不足,本发明提供了一种基于视觉自注意力模型的小样本微调方法,采用在大规模数据集上进行预训练和在小样本任务上进行微调的流程,视觉自注意力模型被用作主干网络,同时构建一个可学习的转换模块norm adapter,由两个向量组成,用于校正原始视觉自注意力模型归一化层的增益和偏置,norm adapter位于视觉自注意力模型ViT的所有归一化层之后,通过逐元素的乘法和加法实现;在预训练期间,使用大规模数据集上以全监督或自监督方式训练的主干网络;在微调过程中,采用原型网络ProtoNet分类头。本发明计算简便,通过逐元素相乘和相加即可实现,因此占用的存储和计算资源比较少,有利于预训练模型投入实际应用场景。

[0006] 本发明解决其技术问题所采用的技术方案包括如下步骤:

[0007] 步骤1:构建主干网络;

[0008] 采用改进的视觉自注意力模型ViT作为主干网络;

[0009] 原始视觉自注意力模型由一个补丁嵌入层和N个transformer层组成;经过补丁嵌入层,输入图像被编码成一定数量的token向量,再与位置编码相加后,输入的token向量连同CLS token被送入N个transformer层中;最终,经过N个transformer层和一个归一化层LayerNorm后,CLS token用于分类或其他目的;每个transformer层包含两个归一化层LayerNorm,一个MLP块和一个多头自注意力块MHSA;

[0010] 构建一个可学习的转换模块,由两个向量组成,用于校正原始视觉自注意力模型归一化层LayerNorm的增益gain和偏置bias,将可学习的转换模块称为norm adapter;所述norm adapter位于视觉自注意力模型ViT的所有归一化层之后,通过逐元素的乘法和加法实现;如公式(1)所示,Scale和Shift分别是norm adapter的两个可学习向量,y是归一化层的输出,⊙代表逐元素乘法:

[0011] $$h = Scale \odot y + Shift \quad (1)$$

[0012] norm adapter的参数Scale和Shift的结构与归一化层的参数增益gain和偏置bias相同,分别被初始化为全1和全0的向量;在微调时,只有参数Scale和Shift被更新,其他参数在预训练后被冻结,不进行优化;

[0013] 步骤2:在预训练期间,使用大规模数据集上以全监督或自监督方式训练的主干网络;

[0014] 步骤3:在微调过程中,采用原型网络ProtoNet分类头;该分类头根据查询图像和原型在嵌入空间中的距离,产生一个概率分布,如公式(2)所示:

[0015] $$p_{\phi}(y=k \mid x)=\frac{\exp(-d(f_{\phi}(x),c_{k}))}{\sum_{k'}\exp(-d(f_{\phi}(x),c_{k'}))} \quad (2)$$

[0016] 其中,f_{Φ}是主干网络,将输入编码到特征空间;c_{k}为类别k的原型,是属于类别k的特征的平均值;d是度量函数;具体来说,各个类别的原型是由支持集中每类样本求均值计算出来的,并将进行数据增强后的支持集作为伪查询集,然后由原型和伪查询集之间的余弦距离计算损失,更新参数;

[0017] 损失函数选择交叉熵损失。

[0018] 优选地,所述自监督方式,采用DINO和MOCO v3算法在ImageNet-1K数据集上训练主干网络;所述全监督方式,主干网络是在ImageNet-21K数据集上训练得到的。

[0019] 优选地,所述度量函数使用余弦距离。

[0020] 本发明的有益效果如下:

[0021] (1)本发明作为一种小样本微调方法,更新的参数量小,仅相当于全部微调(full-tuning)所需更新参数量的0.045%,计算简便,通过逐元素相乘和相加即可实现,因此占用的存储和计算资源比较少,有利于预训练模型投入实际应用场景。

[0022] (2)本发明在real、clipart、sketch、quickdraw四个数据集上的测试结果明显优于全部微调(full-tuning)、bias-tuning,visual prompt tuning等方法。

附图说明

[0023] 图1为视觉自注意力模型ViT的transformer层示意图。

[0024] 图2为加入norm adapter后的transformer层示意图。

具体实施方式

[0025] 下面结合附图和实施例对本发明进一步说明。

[0026] 本发明采用了在大规模数据集上进行预训练和在小样本任务上进行微调的流程，没有在目标域的基类数据上进行训练。视觉自注意力模型(ViT)被用作主干网络，一个普通的视觉自注意力模型由一个补丁嵌入层(patch embedding)和N个transformer层组成。经过补丁嵌入层，输入图像被编码成一定数量的token向量，在与位置编码相加后，输入的token向量连同CLS token被送入N个transformer层中。最终，经过N个transformer层和一个归一化层(LayerNorm)后，CLS token用于分类或其他目的。每个transformer层包含两个归一化层(LayerNorm)，一个MLP块和一个多头自注意力块(MHSA)。图1为视觉自注意力模型(ViT)的transformer层，对应全部微调方法(Full-tuning)，transformer层中的归一化层(LayerNorm)，MLP块和多头自注意力块(MHSA)都是可学习的。

[0027] 本发明提出使用一个可学习的转换模块，由两个向量组成，来校正归一化层(LayerNorm)的增益(gain)和偏置(bias)，称为“norm adapter”。“norm adapter”位于视觉自注意力模型(ViT)的所有归一化层之后，以与增益和偏置相同的方式对激活值进行缩放和移位，具体来说，是通过逐元素的乘法和加法实现的，如公式(1)所示，Scale，Shift分别是“norm adapter”的两个可学习向量，y是归一化层的输出，⊙代表逐元素乘法。

[0028] $$h = Scale \odot y + Shift \quad (1)$$

[0029] “norm adapter”的参数s1和s2的形状与归一化层的增益(gain)和偏置(bias)相同，分别被初始化为全一和全零的向量，因此，与微调前的原始预训练模型相比，带有“norm adapter”的模型在计算结果上没有变化。在微调时，只有“norm adapter”的参数Scale和Shift被更新，其他参数在预训练后被冻结，不进行优化。图2为加入“norm adapter”后的transformer层，对应本发明提出的微调方法，transformer层中只有“norm adapter”的参数Scale、Shift是可学习的。

[0030] 在预训练期间，使用在大规模数据集上以全监督或自监督方式训练的主干网络。对于自监督算法，采用DINO和MOCO v3算法在ImageNet-1K数据集上训练主干网络；对于全监督算法，主干网络是在ImageNet-21K数据集上训练得到的。

[0031] 在微调过程中，采用了原型网络(ProtoNet)分类头。该分类头根据查询图像和原型在嵌入空间中的距离，产生一个概率分布，如公式(2)所示：

[0032] $$p_{\phi}(y=k \mid x)=\frac{\exp(-d(f_{\phi}(x),c_{k}))}{\sum_{k'}\exp(-d(f_{\phi}(x),c_{k'}))} \quad (2)$$

[0033] f_{ϕ}是主干网络，将输入编码到特征空间。c_{k}为类别k的原型，是属于类别k的特征的平均值。d是度量函数，这里使用的是余弦距离。具体来说，原型是由支持集计算出来的，并将数据增强后的支持集作为伪查询集。然后由原型和伪查询集之间的余弦距离计算损失，更新参数。损失函数选择交叉熵损失(Cross Entropy)。

[0034]

[0035] 本发明采用视觉自注意力模型(ViT)作为主干网络，包括ViT-Base/16和ViT-Small/16，对于ViT-Base/16，我们分别采用监督学习方法在ImageNet-21K数据集上训练，

采用MOCO-v3算法在ImageNet-1K数据集上训练得到预训练主干网络;对于ViT-Small/16,采用DINO算法在ImageNet-1K数据集上训练。

[0036] 在下游任务上微调和评估时采用了real、clipart、sketch、quickdraw四个数据集,它们是DomainNet的子数据集,包含相同的类别名。

[0037] 在微调和评估过程中,采用30-way 5-shot的形式来构建小样本任务,每个任务包含5个类别的数据,每类数据有5张带标注样本和15张查询样本;所有图像均被调整成224*224分辨率大小;用于生成伪查询集的随机数据增强包括颜色抖动、水平翻转和平移;微调过程中有三个超参数比较关键:学习率、迭代次数和优化器,由于每个任务中的样本有限,最终性能对超参数的选择比较敏感,所以对于各种情况,根据验证集上50个任务的平均准确率来选择超参数,优化器从Adam或SGD中选择,学习率和迭代次数从经验范围内选择,分别为[1e-1,1e-2,1e-3,1e-4,1e-5,1e-6]和[20,50,80,100];最后,从测试集中随机选取600个任务进行评估,计算平均精度作为最终结果。所有的实验均采用固定的随机数种子。

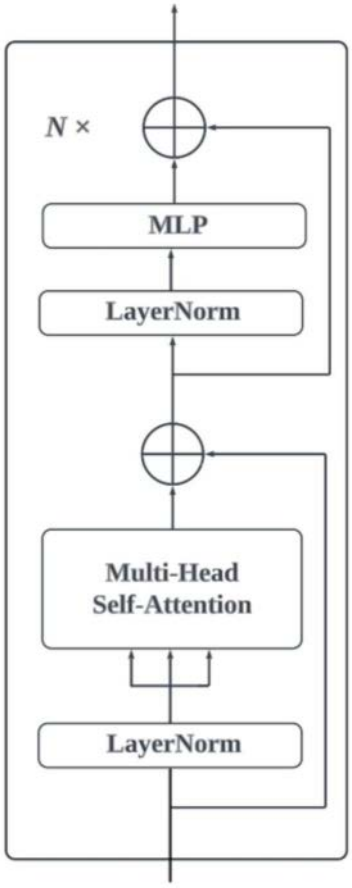


图1

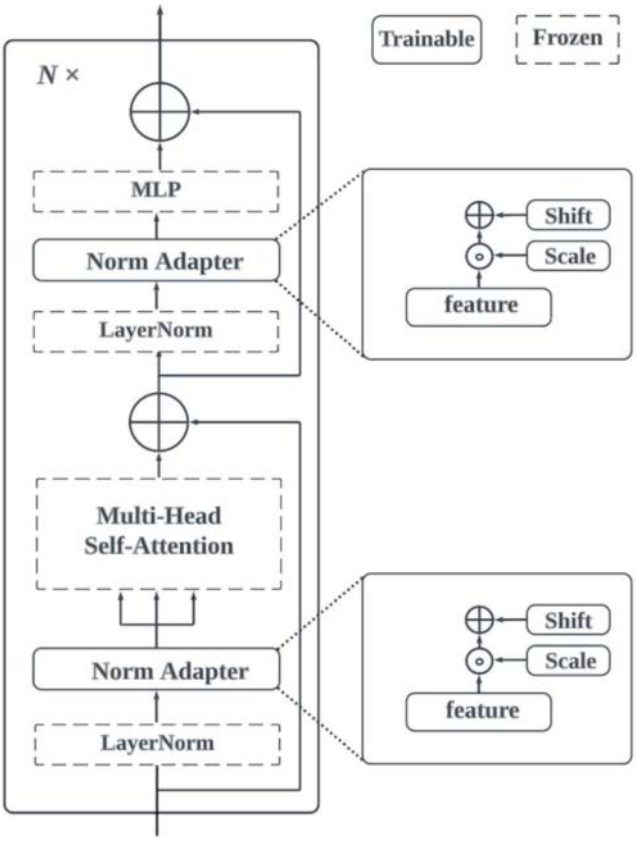

Trainable
Frozen
N ×
MLP
Norm Adapter
LayerNorm
Multi-Head
Self-Attention
Norm Adapter
LayerNorm
Shift
Scale
feature
Shift
Scale
feature

图2