烟火识别任务大作业

刘芳甫
清华大学
liuff19@mails.tsinghua.edu.cn

李泽润
清华大学
lizr19@mails.tsinghua.edu.cn

摘要

烟火检测任务是在监控视频或者图像中进行烟火图像分类或者定位，在消防领域具有独特的意义。本文基于当前深度学习和计算机视觉的背景下来实现该任务。由于自己收集的训练集较少，所以借鉴了迁移学习（Transfer Learning）的思想对该问题进行实现，本文首先采取了Resnet作为Backbone对该识别问题进行 finetune，取得了较好的效果。基于当前自监督学习（Self-Supervised Learning）的流行，并且考虑该任务或者扩展任务能更好地具有泛化性，本文又引入MAE（Masked AutoEncoder）的方法。出于对当前火热模型的实用性探索和一定的好奇，我们又使用了DenseNet和（VIT）Vision Transformer的模型对该下游问题进行实现。尽管引入这类大模型有些大材小用，本文希望不仅是基于此任务，而是能通过这些方法启发我们未来更多的思考和工作，锻炼并提升自己的能力，不足之处希望老师和助教建议指正。

1. 背景介绍

烟火检测任务是在监控视频或者图像中进行烟火图像分类或定位，在消防安全领域具有独特的意义。常见的烟雾传感器是靠检测物质燃烧后空气中浓度升高的二氧化锡等来报警，而视频监控中基于视觉的烟火检测可以覆盖较为广阔的区域，比如街道上的火灾检测、无人机森林防火巡查等。本文基于深度网络和当前流行的深度网络模型对该问题进行实现。

本文采用自己收集数据集作为训练集的方法，利用迁移学习的知识来对我们的任务进行模型的微调，然后用 BoWFire 的数据集来作为我们的测试集进行方法上的评估。

实际中，基本没有人会从零开始（随机初始化）训练一个完整的网络框架，因为相对于网络而言，很难得到一个足够大的数据集（网络很深， 需要足够大数据集）。而且收到计算资源的限制。所以通常的做法是利用在很大一个数据集上进行预训练得到一个网络，然后将这个网络的参数作为目标任务的初始化参数或者固定这些参数。

出于对当前火热模型的探索和好奇，本文使用了较多当前主流的 Image Model 分别对该任务进行了实现，并比较性能。 本文应用的主要方法如下：

（1） **微调 ResNet**，使用预训练好的 resnet34 模型初始化自己的网络，而不是随机初始化，然后在这个烟火检测的任务下进行微调。

（2） 考虑到现在**自监督模型**的热度，尽管将该模型用于此会大材小用，但作者希望能从该方法中获得启发并得到锻炼。我们采用 Kaiming He 近段时间提出的 **MAE** (Masked AutoEncoder) 自监督学习的方法进行任务实现。**将其中的 Encoder 看作是固定的特征提取器**，最后加一个 MLP 网络实现烟火识别任务。

（3） **微调 VIT 和 Dense,** 使用预训练好的 VIT 和 DenseNet 模型初始化自己的网络，而不是随机初始化，然后在这个烟火检测的任务下进行微调。

上述方法中，**ResNet** 取得了较好的表现，在准确率上**高达 91%**。在自监督模型较火的基础上，我们也自己修改了 **MAE** 的代码让其适用于我们的任务，采用 MAE 作为固定的特征提取器的方法，尽管准确率有所下降，只能达到 81%，但是考虑到我们自行选取的烟火数据集可能和预训练的数据上有 domain 上的偏差，如果预训练模型的数据有大量无标注的烟火数据，再用较少有标注的烟火数据做下游任务，效果将会非常好，会优于有监督的训练。后面出于探索现有火热模型在下游任务上的能力，我们又使用了 **DenseNet** 和 **VIT** 对其

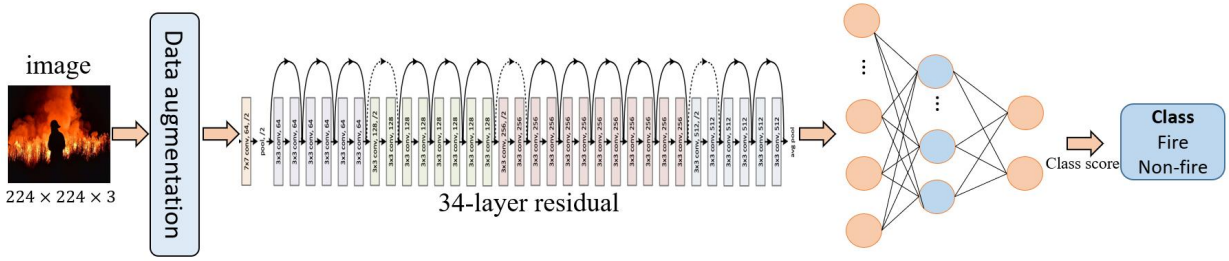


figure 1 基于ResNet 迁移学习的Pipeline

识别任务进行实现，特别的 **VIT 的方法**在准确率上能**达到 92%**。

2. 相关工作

BoWFire method [1]：该方法是本文采用的评估测试集，该方法是在 2015 年提出，该方法较为传统，是基于颜色特征的分类，结合super pixel区域的纹理分类。该方法能有效的减少假阳性，并使用较少的参数。

残差神经网络（ResNet）[2]：残差神经网络（residual neural network）是由微软亚洲研究院的Kaiming He等人提出的。其主要贡献是发现了“退化现象”（Degradation）并针对退化现象发明了Shortcut connection，极大的消除了深度过大的神经网络训练困难的问题。并且由于resnet的强大和多任务的适用性，现有很多工程上的任务都会采用ResNet作为backbone来实现。

DenseNet [3]：相比ResNet，DenseNet提出了一个更激进的密集连接机制：即互相连接所有的层，具体来说就是每个层都会接受其前面所有层作为其额外的输入。相比ResNet，这是一种密集连接。而且DenseNet是直接concat来自不同层的特征图，这可以实现特征重用，提升效率，这一特点是DenseNet与ResNet最主要的区别。

AutoEncoder(AE)：自编码机是一种无监督的神经网络模型，它可以学习到输入数据的隐含特征，这称为编码（encoding），同时用学习到的新特征可以重构出原始的输入数据，称之为解码（decoding）。自动编码机学习到的新特征可以送入到有监督的学习模型中，所以自编码机可以起到特征提取器的作用。

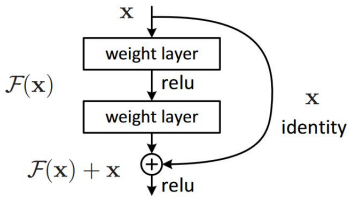
MAE(Masked AutoEnCoder) [4]：MAE 是一种自编码方法，给定原始信号的部分观测，然后对原始信号进行重建。和其他自编码器的方法类似，MAE中一个encoder将一个观测信号映射称为隐表示，一个decoder使用隐表示对原始信号进行重建。与传统的AE不同，MAE采用了非对称的设计，允许encoder只处理部分，观测到的信号，并用一个轻量级的decoder用隐表示和mask token进行重建。本扩展方法可以学习大容量的模型并更好的泛化，下游任务的迁移性优于有监督的训练，并显示出非常好的扩展行为。

ViT(Vision Transformer) [5]：虽然Transformer体系结构已经成为自然语言处理任务的事实标准，但其在计算机视觉中的应用仍然有限。在视觉中，注意力要么与卷积网络结合使用，要么用于替换卷积网络的某些组件，同时保持其整体结构。实际上，这种对CNN的依赖是不必要的，直接应用于图像块序列的纯变换器可以很好地执行图像分类任务。同时训练所需的计算资源也大大减少。

3. 方案设计

3.1. 基于ResNet的烟火识别任务的迁移学习

ResNet的主流网络框架主要基于解决神经网络随着深度增加的degradation problem。所以采取了如下的Residual learning frame:



因为ResNet作为backbone的表现无论在任何任务上都表现的非常好，所以我们给出如下几种有效性解释：

（1） 使得网络更容易在某些层学到恒等变换（identity mapping）。在某些层执行恒等变换是一种构造性解，使更深的模型的性能至少不低于较浅的模型。

（2） 可以利用模型集成的思想（ensemble），残差网络可以看成是很多网络集成，类似现在的平均老师模型（mean teacher）的一种思想。

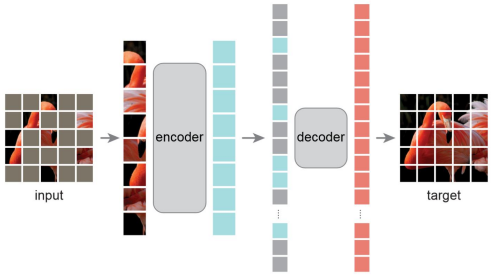
（3） 残差网络使得信息更容易在各层之间流动，包括在前向传播的时候提供特征复用，在反向传播的时候缓解梯度消失。

基于上述分析，我们选取了ResNet34 作为我们的backbone。其pipeline如图表 1 所示，我们首先对输入的烟火图像数据做了data augmentation，然后输入到已经预先训练好的残差网络中，最后接一个mlp完成我们最后的识别任务，通过相应的class score给出我们的预测结果即可。ResNet适用的任务非常之广，实际上烟火检测任务只是下游任务中一个非常简单的任务，所以通过实践，我们发现ResNet的backbone在此任务上的迁移效果较佳。

3.2. 基于自监督的MAE的烟火特征学习方法

解决该分类问题实际上并用不到这样的SSL模型，的确上是大材小用，但是作者希望通过这样的机会进行当前流行模型的下游任务实现，既是希望对比有监督的方法，同时也希望通过该任务迁移看到网络更多的可能性，在锻炼自己能力的同时也对相应的实验结果和网络框架有更多的思考。

关于MAE的核心思想如下图所示：



Masked Autoencoder使用了掩码机制，利用编码器将像素信息映射为语义空间中的特征向量，而使用解码器重构原始空间中的像素。MAE使用的是非对称的Encoder-Decoder架构，即编码器只能看到未被遮蔽的部分像素块信息，以节省计算开销，而解码器解码的是所有像素块的特征信息。

这样的一种Representation Learning 相对于有监督学习来说，从理论上更能学习到non-trivial的特征，并且节省了人工大量标注数据的成本，同时参数量相对于大型的CNN网络来说也大大减少，所以自监督学习在海量数据的大背景下有着较好的发展前景。

我们主要在原有的MAE网络框架的基础上，将其修改，适用于我们的烟火识别任务。我们希望的是能借助自监督的方法，学到烟火图像中一些更non-trivial的特征，从而能泛化到更多的烟火场景中，并提高模型的鲁棒性。

我们首先基于MAE预训练好的参数，然后将其Deocoder部分去掉，我们利用MAE的encoder部分作为烟火图像的特征提取器，将其得到的特征中的class token接入我们的MLP head中，然后进行分类预测。器pipeline如图 2 所示。

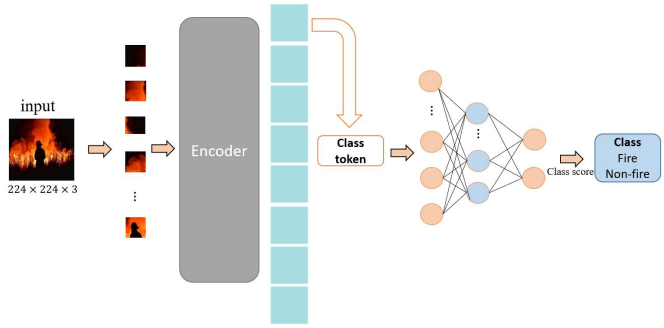
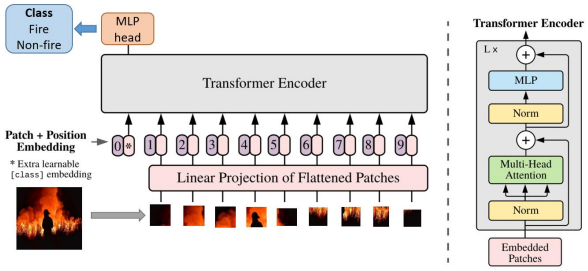


figure 2 基于MAE的烟火检测pipeline

3.3. 基于ViT的烟火检测实现

由于只是出于对当前流行模型ViT在下游任务上的表现的好奇和兴趣，所以我们并没有像MAE和ResNet那样修改ViT网络具体的一些框架来对任务进行实现，我们只是使用了已经预训练好的ViT模型，然后在我们的任务上微调。具体的pipeline如下图所示：



令人惊喜的是，它在我们测试的数据集BoWFire上取得了非常好的表现，甚至高于ResNet一个百分比，**达到 92%的准确率**。可能是通过这样Transformer的形式能提取出烟火图像中更适用的特征，这也给了我们一定的启发。由于我们的计算资源有限，无法在相应的大量的烟火数据集上进行训练，所以进一步的猜想还需要未来计算资源的支持。

3.4. 基于DenseNet的烟火检测实现

同样出于好奇，我们在尝试完用ResNet作为Backbone进行迁移任务的微调之后，我们又尝试了以DenseNet作为Backbone来进行该下游任务的实现。

DenseNet脱离了加深网络层数(ResNet)和加宽网络结构(Inception)来提升网络性能的定式思维,从特征的角度考虑,通过特征重用和旁路(Bypass)设置,既大幅度减少了网络的参数量,又在一定程度上缓解了

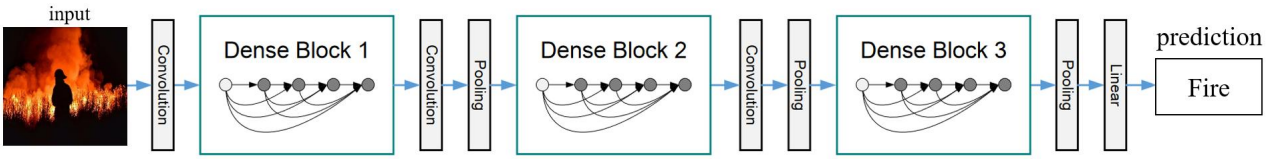
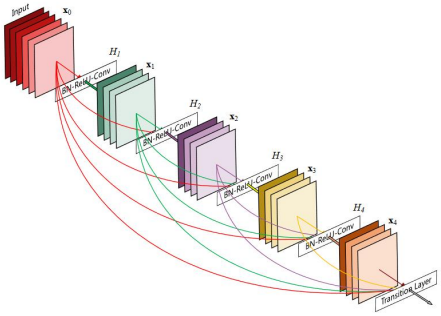


figure 3 基于DenseNet 的烟火检测网络

gradient vanishing问题的产生，结合了信息流和特征复用。核心思想如下图所示：



同样因为计算资源和训练集有限的问题，我们只能在原先DenseNet的基础上去做Finetune，具体的网络框架如图 3 所示。最终取得的效果和ResNet的是可比的，这是因为该下游分类任务较为简单，不能体现出DenseNet的优势，实际上在工业中，DenseNet使用的相较于ResNet还是偏少，因为DenseNet的网络设计层面上更为复杂，参数也更难调整。

4. 实验结果

4.1. 数据集

为了能完成上述的任务，我们自行在网上收集了部分的数据集作为训练，其中包括 **845 张正样本**和 **748 张负样本**，实际上这个收集的数据集非常小。测试集是课程文件中所给的BoWFire数据集（包括 226 张图片，其中 119 张图片是fire，107 张图片是without fire）。**(注：在训练过程中，我们未用到任何关于BoWFire的数据集信息)**，尽管训练的数据集较少，但是我们通过上述方法中的网络设计，都能得到较好的结果。

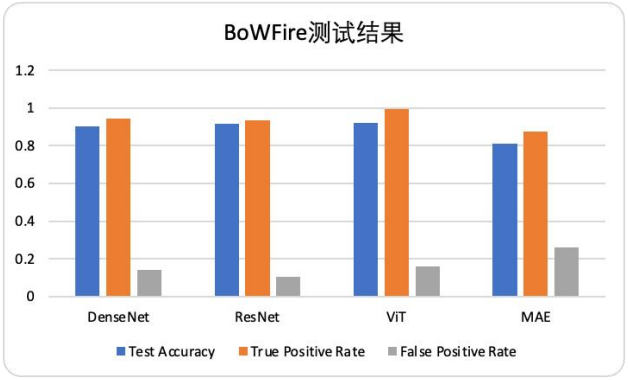
4.2. 训练过程和实验结果

训练过程我们首先对输入的训练烟火图像采取了数据增强（data augmentation），batch size默认选取是 4，在训练过程中我们采用固定步长衰减（在固定的训练周期后，以指定的频率进行衰减），优化器选取SGD。计算资源是一张RTX 3060 显卡。

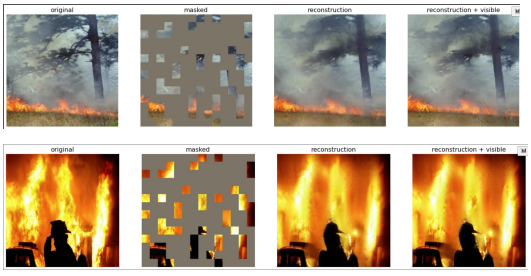
最终得到的结果如下所示：

Model	准确率	真阳率	假阳率
基于ResNet34	**91.72%**	95.01%	**10.02%**
基于DenseNet	90.56%	96.02%	12.10%
基于ViT	**93.08%**	**99.26%**	14.21%
基于MAE	82.09%	89.01%	18.01%

可视化后效果如下：



可以看到我们的方法在该任务上的效果都较好，但是MAE的方法相对于其他模型来说准确率和真假阳性都较低，我们猜测可能是我们所用的MAE的预训练模型的预训练数据与我们测试的数据集domain上有偏差。但是我们可以通过如下的实验看到MAE学到的特征和信息重建的能力：



其实MAE的能力很强，这是在大规模数据集上较为成功的办法，实际上我们的烟火检测任务非常简单，并不需要动用如此的模型，但是出于探索和好奇的心理，作者还是尝试使用了这种办法作为我们的实践方法之一。

从上述的实验结果看到，ViT，ResNet，DenseNet作为Backbone的效果都非常好。由于实践时间，计算资源和数据集等多方面的限制，我们并不能验证更多的猜想，目前只能基于任务给出我们尝试的实现。

5. 结果讨论

总体来说，我们在该下游任务上的实践效果较好，取得了不错的表现。

在准确率和真假阳性的结果中，我们进一步对实验现象进行了分析，我们在实验过程中发现，有一些假阳性的识别错误实际上是一些颜色和颜色区域大小非常接近火焰的图片（比如**夕阳图片**）。

对此我们希望在未来如果有机会能够解决这个问题。夕阳的图片实际上是**hard negative**，针对这类的问题（并不仅限于此烟火检测任务）我们给出如下几个可行性方案（以后如果有时间和计算资源将加以实现验证）：

（1） **Mixup:** 这是FAIR提出的一种数据增强方法：利用两张不同的图像随机线性组合，而同时生成线性组合的标签（soft label）。比如我们可以用 0.9*烟火+0.1*夕阳来对此实现。

（2） **度量学习/对比学习：**度量学习是从数据中学习一种度量数据对象间距离的方法，其目标是让学得的距离度量下，类内距离小，类间距离大，其实这样能帮助我们区分开类似这种hard negative的干扰。

（3） **在深度学习中引入传统方法监督，**比如BoWFire中涉及到的super pixel的概念，用特征颜色和区块范围作为进一步的监督信息，而且火焰一般是有明显的边缘信息的，我们可以在深度网络训练的过程中引入梯度算子等方法。

最后，尽管这个烟火检测的任务是一个较为简单的任务，但是给了我们实践实操的机会，本文采用上述的方法尽管有时候有些大材小用，但是作者希望借此机会进一步锻炼自己的能力，并对当前主流框架有进一步的认知和了解，从而有所思考有所启发。最后感谢老师和助教给了我们实践这次大作业的机会，我们从中收获颇丰。

References

[1] Chino, Daniel YT, et al. "Bowfire: detection of fire in still images by integrating pixel color and texture analysis." 2015 28th SIBGRAPI conference on graphics, patterns and images. IEEE, 2015.

[2] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[3] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[4] He, Kaiming, et al. "Masked autoencoders are scalable vision learners." arXiv preprint arXiv:2111.06377 (2021).

[5] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).