

# 基于卷积神经网络的吸烟打电话识别及网络可视化

刘知安

[csliu zhian@mail.scut.edu.cn](mailto:csliu zhian@mail.scut.edu.cn)

**摘要：**在某些重要的场合，对人物的行为作出特定限制十分有必要，比如加油站禁止吸烟，驾驶员禁止打电话，博物馆禁止拍照等。本文以识别吸烟和打电话的行为为目标，基于最先进的的卷积神经网络（CNNs）模型，针对监控视频中的图像数据，分别基于 VGG 和 ResNet 网络设计了用于吸烟打电话检测的 SCVGG 和 SCResNet 模型，在 Top-1 准确率和 mAP 两个指标上均取得了较高分数，并于 2020 中国华录杯吸烟打电话赛道 A 榜取得 99.62 的分数。此外，为了验证 CNN 强大的特征提取能力以及网络着重关注的区域，基于导向反向传播和 Grad-CAM 技术，本文对所构建的网络模型进行了可视化，可视化结果表明，网络会着重关注图像中的烟头和手持电话区域，和人为判别的准则十分相似。本项目源代码已开源至 <http://github.com/LiUzHiAn/MLproj>。

**关键词：**图像分类，深度学习，卷积神经网络，神经网络可视化

## 1. 介绍

行为规范，即指某些特定的场景会对人物的行为做出特定的限制，比如加油站禁止吸烟，驾驶员禁止打电话，博物馆禁止拍照等。随着计算机人工智能的发展，这些禁止行为或者不文明行为都可通过基于视频监控的行为检测算法进行监控、发现以及适当时给与警告。本文将吸烟打电话识别视为一个图像分类问题。由于监控视频中的图像清晰度有限，而且人物的像素大小不一，使用传统方法的特征提取方法（例如 HOG<sup>[13]</sup>、SIFT<sup>[8]</sup>等）往往效果不佳，从而对后续的分类产生影响。自从 2012 年以来，卷积神经网络（CNNs）已经主宰了图像分类领域，ImageNet<sup>[2]</sup>的出现也催生了许多新颖的网络模型，这也使得快速实现其他分类任务成为可能。

然而，虽然很多模型在 ImageNet 上的分类准确度都超过了人类的判别水平，但对于网络中各模块的具体功能我们还尚未完全清楚，缺乏可解释性<sup>[3]</sup>。有许多研究者已经开始研究深度网络可解释性相关的问题，尝试探索为什么模型将一张图片判别为某个具体的类。导向反向传播<sup>[12]</sup>（Guided-backpropagation，GBP）让我们可以“看到”网络的内部感兴趣的区域，而借助 Grad-CAM<sup>[10]</sup>，我们可以解释网络的分类判别情况。

本文的主要工作如下：

- 首先，基于最先进的卷积神经网络结构，设计了一种用于判别吸烟打电话行为的模型，在 Top-1 准确率和 mAP 两个指标上均取得了较高分数；
- 其次，基于 GBP 和 Grad-CAM 算法，本文对所构建的网络模型内部进行了可视化，实验结果表明，网络内部关注的区域和人为图像分类的判别理念十分相似，也验证了本文方法的有效性。

## 2. 相关工作

在(2.1)小节，我们对一些经典且强大的图像分类网络模型进行了概括。在(2.2)小节，我们将对网络可视化中的导向传播算法和 Grad-CAM 算法进行介绍。

### 2.1 图像分类网络模型

AlexNet<sup>[7]</sup>、VGG<sup>[11]</sup>、ResNet<sup>[5]</sup>等神经网络模型都是很成功且影响巨大的网络，它们在 ILSVRC 竞赛中都曾取得了很好的成绩，也为后续的神经网络结构搜索打下了基础，表(1)给出了三种网络架构在 ImageNet 上的 Top-1 和 Top-5 准确度。

网络模型	Top-1 Acc	Top-5 Acc	参数量
AlexNet	63.3%	84.6%	60M
VGG-16	74.4%	91.9%	138M
VGG-19	74.5%	92.0%	144M
ResNet-50	77.2%	93.3%	22M
ResNet-101	78.3%	94.0%	25M
ResNet-152	78.6%	94.3%	60M

表 1. AlexNet、VGG 和 ResNet 网络在 ImageNet 上的 Top-1 和 Top-5 准确度

AlexNet 是计算机视觉中神经网络结构的先驱，AlexNet 中包含 5 个卷积层和 3 个全连接层，每个卷积层和全连接层后都使用了 ReLU 作为激活函数，为防止过拟合，网络在全连接层中使用了 Dropout，且在训练的过程中使用了水平翻转、随机裁剪等数据增广策略，提升了模型的分类准确率和鲁棒性。

VGGNet 是由 Visual Geometry Group 提出的，该网络说明了通过增加网络的深度可以提高模型的性能。该模型的主要特点是，在卷积层中使用连续的 3\*3 卷积替代 AlexNet 中的大卷积核（例如 11\*11 和 7\*7 卷积核），池化层的池化大小也都为 2\*2，VGGNet 中提出了 VGG-16 和 VGG-19 两种网络结构，前者包含 13 个隐藏层和 3 个全连接层，后者则包含 16 个隐藏层和 3 个全连接

层。VGGNet 的缺陷在于计算量较大（140M 参数）且训练需要较大的显存。

ResNet 是最成功的网络结构之一，在训练深度神经网络时，梯度消失或梯度爆炸是一个严重影响模型性能的因素，而 ResNet 中视之为一个优化问题，作者提出了以下残差函数

$$H(x) = F(x) + x$$

并认为理想情况下的  $H(x)$  总是和  $x$  很接近，不要让整个网络去直接学习  $H(x)$ ，而是学习残差量  $F(x)$ ，进而提出了残差连接的概念，图(1)是 ResNet 中残差块和普通卷积神经网络模块的对比。

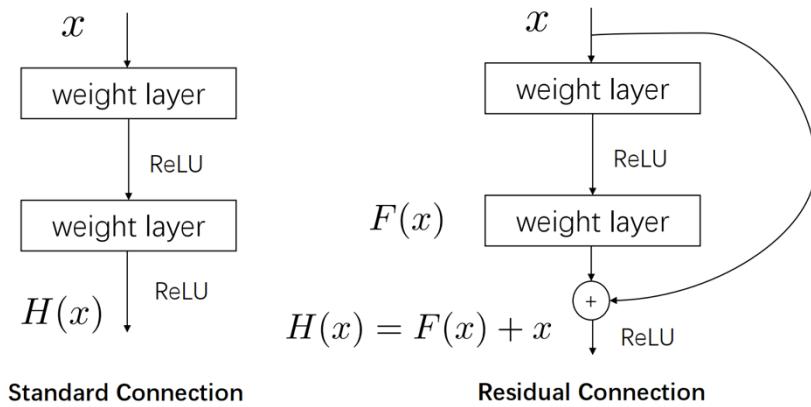


图 1. 普通连接 V.S. 残差连接，网络学习残差量  $F(x)$  而不是直接学习  $H(x)$

ResNet 的提出也使得训练大型深度神经网络成为可能，作者分别实验了 Res18、Res34、Res50、Res101 和 Res152 五种规模的网络，更深的网络效果更好，也进一步验证了 VGGNet 中的猜想。在网络中的每个卷积层后都使用了 Batch Normalization 和 ReLU 作为激活函数。

## 2.2 网络可视化

在可视化网络时，有两个非常重要的问题，1) 网络具体在关注那些东西？2) 为什么网络将一张图片判别为某个具体的类？这两个问题可以分别通过导向反向传播和 Grad-CAM 来回答。

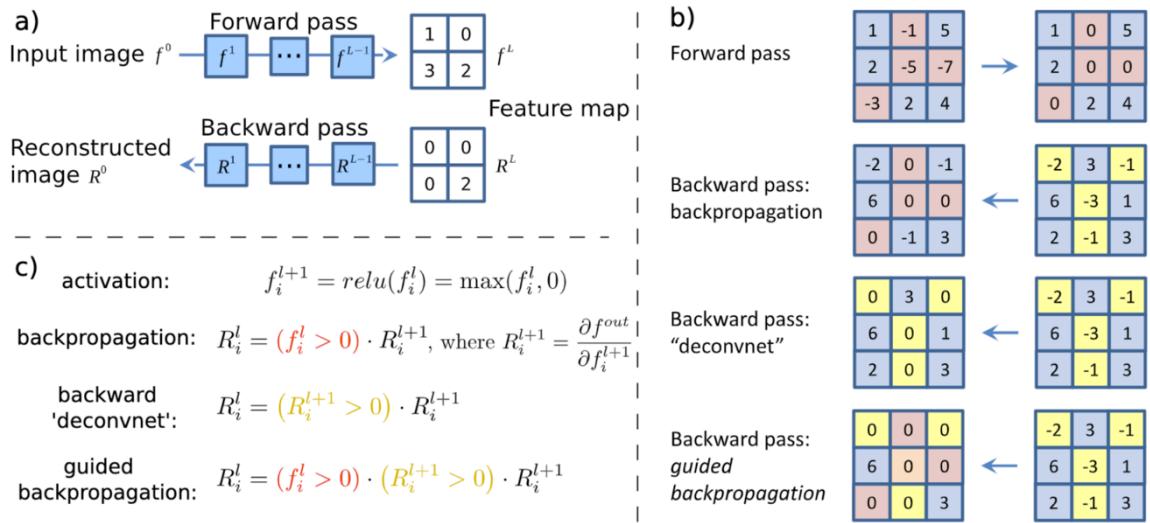


图 2. 向前传播、向后传播、反卷积和导向反向传播对比

在使用 ReLU 作为激活函数的前提下，图(2)对比展示了向前传播、向后传播、反卷积和导向反向传播的不同点。在网络反向传播时，只会传播值为正数的梯度；反卷积则是对梯度值进行激活；而导向反向传播（Guided-backpropagation, GBP）综合考虑了反卷积（DeConv）和反向传播过程。记  $f^{out}$  为网络的输出， $f^l$  为向前传播时第  $l$  层的输出， $R^l$  为第  $l$  层反向传播时的梯度，通过求解

$$R_i^l = (f_i^l > 0) \cdot (R_i^{l+1} > 0) \cdot R_i^{l+1}$$

我们可以得到网络在第  $l$  层所关注的地方，从而对网络的内部展开可视化解释。

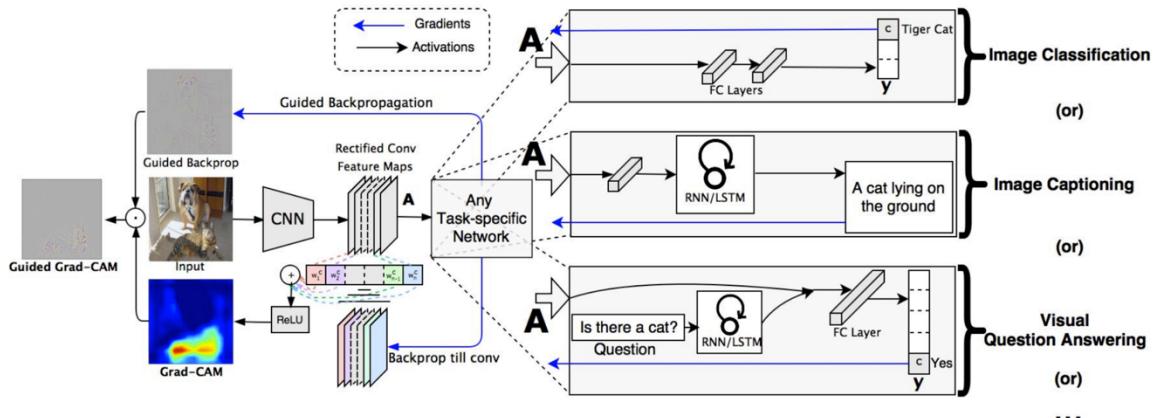


图 3. Grad-CAM 计算示意图，图片来源于<sup>[10]</sup>

Grad-CAM 是类激活图（Class Activation Map, CAM）技术中的一种，它表示的是输出中某个类别对输入图像各区域的敏感程度。对于猫狗分类问题，

CAM 可以很直观地告诉我们，模型是因为看到了图像中的哪个部分才将图像判别为猫或狗的。Grad-CAM 可以用在很多不同的任务中，其计算流程如图(3)所示。以图像分类任务为例，记  $y_c$  为网络对类别 c 的预测得分（在 Softmax 计算之前）， $A_{ij}^k$  为网络最后一个特征图第 k 个通道上第 i 行第 j 列处的值。通过

反向传播算法，计算得到预测得分关于特征图上每个位置处的梯度  $\frac{\partial y_c}{\partial A_{ij}^k}$ ；

随后对每个通道求得全局平均  $\alpha_k^c = \frac{1}{\text{num of points}} \sum_i \sum_j \frac{\partial y_c}{\partial A_{ij}^k}$ ，即网络对第 k 个通道的敏感程度；然后将  $\alpha_k^c$  与特征图  $A^k$  进行线组合得到网络关于类别 c 的二维图  $L_{Grad-CAM}^c = \text{ReLU}(\sum_k \alpha_k^c)$ ；最后将二维图放大到原输入图像的大小即为类别 c 的类激活图。

### 3. 技术路线

#### 3.1. 数据集

本项目的数据集来源于 2020 中国华录杯数据湖算法大赛的吸烟打电话定向算法赛\*，数据集中包含了训练集和测试集，其中训练集中包含了 6150 张图片样本，训练集中的图片包含吸烟、打电话和正常 3 类，本文对 6150 张训练图片划分成 5000 张图片样本的训练集和 1150 张图片样本的验证集。训练集中的图片测试集包含 A 榜和 B 榜，分别包含 1500 张和 3000 张图片，本项目在验证集上对模型进行评估，并公布测试集 A 榜数据的测试结果。图片来源于监控摄像头和网络，难点在于数据集中人物清晰度有限，而且人物的像素大小不一，图(4)展示了吸烟、打电话以及正常图片的示例。



图 4. 吸烟、打电话及正常类别图片示例

#### 3.2. 解决方案及应用场景

由于数据来源不一且人物清晰度有限，运用传统的基于手工提取特征的方法难度较大，基于深度卷积神经网络的强大自适应特征提取能力，本文也采用

---

\* 在此感谢主办方提供的数据

CNN 来对图像进行分类。我们将原问题视为一个分类问题，即吸烟、打电话和正常三个类别，通过输入足量的图片，模型可以进行端到端的训练。在模型训练完毕之后，模型可以对任意测试图片进行判别，对于吸烟、打电话和正常都给出一个分数，通过给定一个阈值，我们可以实时地对某个类别作出及时的报警。训练好的模型通过调优也可以部署到下游的端设备上，在一些特殊的场合下具有广泛的应用场景。

## 4. 方法

### 4.1. 网络模型

我们分别使用 VGG 和 ResNet 的 backbone 作为特征提取器，并将原模型的最后几层替换成只包含一层全连接和 Softmax 层的结构，如图(5)所示。

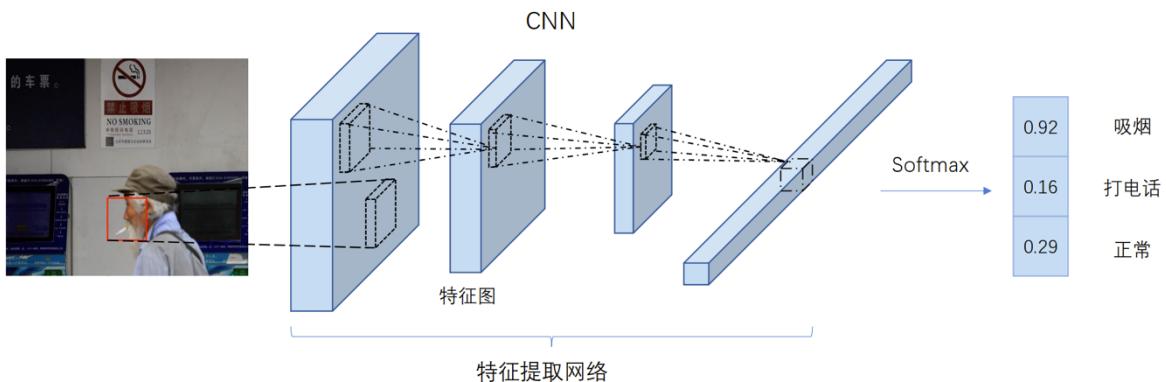


图 5. 网络整体架构图

为了公平对比，我们将 VGG-19 原有的最后三层全连接层替换成只包含一层神经元为 3 的全连接层，记为 SCVGG-17，并作为本文的 baseline。在<sup>[1]</sup>中指出网络越深表达能力越强，效果自然也越好，本文对也参考 ResNet-50、ResNet-101 和 ResNet-151 模型结构，将最后一层全连接替换成输出神经元为 3 的结构，下文中分别记为 SCResNet-50、SCResNet-101 和 SCResNet-151，(5.2)节中的实验结果也验证了这一说法。

我们通过交叉熵来计算真实标签和网络输出之间的损失函数，记  $x$  为网络倒数第二层全连接层的输出，于是损失函数  $L(x)$  可以定义为：

$$L(x) = - \sum_i \log(S(x)_i) y_i, \text{ where } S(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

从而使得网络最大化正确类别的输出得分，同时最小化错误类别的输出得分。

### 4.2. 实验细节

本文使用 PyTorch<sup>[9]</sup>框架来实现算法，用 Adam<sup>[6]</sup>优化器来优化损失函数，

学习率初始化为 0.001，并使用余弦退火策略来改变学习率的大小，周期设置为 80 轮。和 ResNet 中一样，网络输入图片的大小为 224\*224，且对原始的输入图片进行了 mean-std 归一化处理。我们也对 5K 张训练图片进行了数据增广，在每个 batch 中，我们首先对图片的饱和度和色彩进行了微调，二者的变化浮动均设置为 0.05，随后对微调后的结果进行随机水平翻转，最后对水平翻转后的图片进行随机旋转 20 度。Mini-batch 的大小设置为 32，并在训练集上训练了 80 轮，使用的 GPU 为 NVIDIA 2060s 显卡。

## 5. 对比实验

在深度学习模型中，非常关键的两大部分就是模型和数据，模型高效性直接决定了算法性能的天花板，而数据量的多少直接决定了模型的学习收敛情况，在本节中，我们对不同的模型和是否进行数据增广进行了对比实验。

### 5.2.1 不同模型架构对比

表(2)给出了 SCVGG-17、SCResNet-50、SCResNet-101 和 SCResNet-152 四个模型在验证集上的 Top-1 准确度以及模型的参数量，对比结果可以发现，基于 VGG 架构的网络准确度相对偏低，只有 95.00%，而基于 ResNet 架构的网络准确度则均在 98.91% 以上，且随着网络的加深，模型的效果的确更好，说明更深的网络特征提取能力更强，说明残差网络中的残差连接的确发挥了作用。

网络模型	Top-1 准确度	模型参数
SCVGG-17	95.00%	~ 20M
SCResNet-50	98.91%	~ 24M
SCResNet-101	99.06%	~ 42M
SCResNet-152	99.84%	~ 58M

表 2. 不同模型在验证集上的 Top-1 准确度和参数量

图(6)给出了不同网络模型在吸烟、打电话、正常 3 个类别上的 Precision-Recall 曲线，其中红色、蓝色和绿色分别代表 SCVGG-17、SCResNet-50 和 SCResNet-101 模型，而实线、短虚线和虚线则分别代表了正常、打电话和吸烟的 AP 曲线。从曲线图可以发现，SCVGG17、SCResNet-50、SCResNet-101 三者分类准确度依次增加，这和我们在 Top-1 准确度评价指标中得到的结论一致。

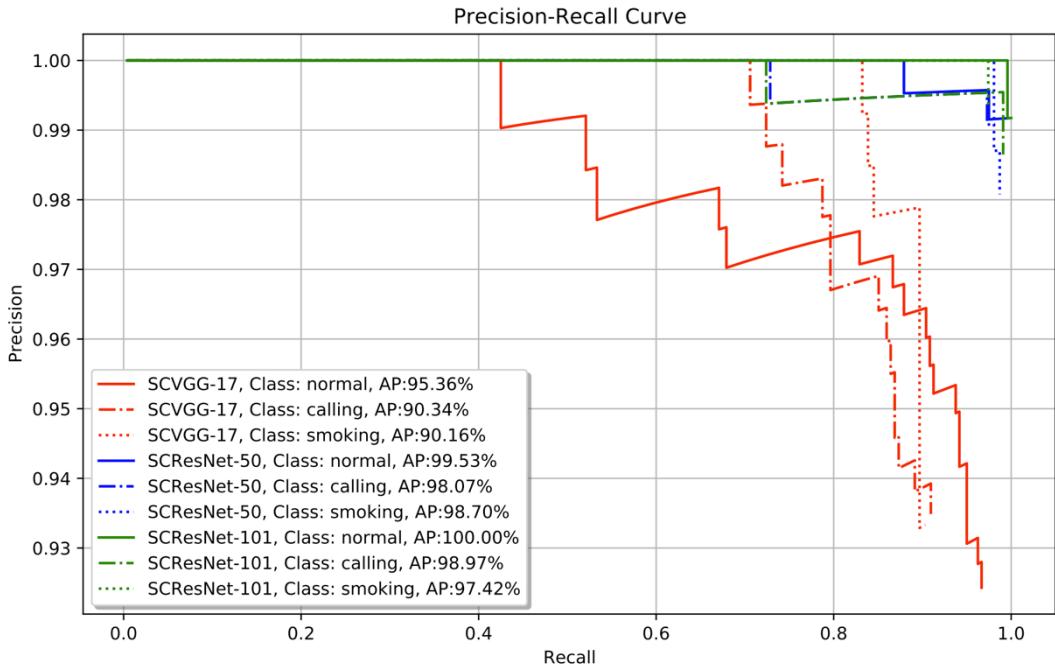


图 6. 不同网络模型各类别 Precision-Recall 曲线

表(3)给出了不同模型各类别的具体 AP 值。从结果上来看，三种方法在 normal 类上的准确率相对吸烟和打电话类别都略高一些，而 SCResNet-50、SCResNet-101 两种网络整体的 AP 值都很高，在 98%以上。值得注意的是，SCResNet-101 在验证集 normal 类别上的 AP 值为 100%，但在吸烟这个类别上的 AP 值相较 SCResNet-50 则下降了 1.68%，这很可能是因为网络对 normal 类别的关注度过分集中，导致其他类别的判别准确度受了影响。

网络模型	AP_s	AP_c	AP_n	mAP
SCVGG-17	90.15%	90.34%	95.36%	91.95%
SCResNet-50	98.70%	98.06%	99.52%	98.76%
SCResNet-101	97.41%	98.97%	99.99%	98.79%

表 3. 不同模型各类别的 AP 值。其中 AP\_c 代表 calling 打电话，AP\_s 代表 smoking 吸烟，AP\_n 代表 normal 正常

### 5.2.2 是否使用数据增广对比

网络模型	Top-1 准确度
SCResNet-50 w/ aug	98.91%
SCResNet-50 w/o aug	94.38%

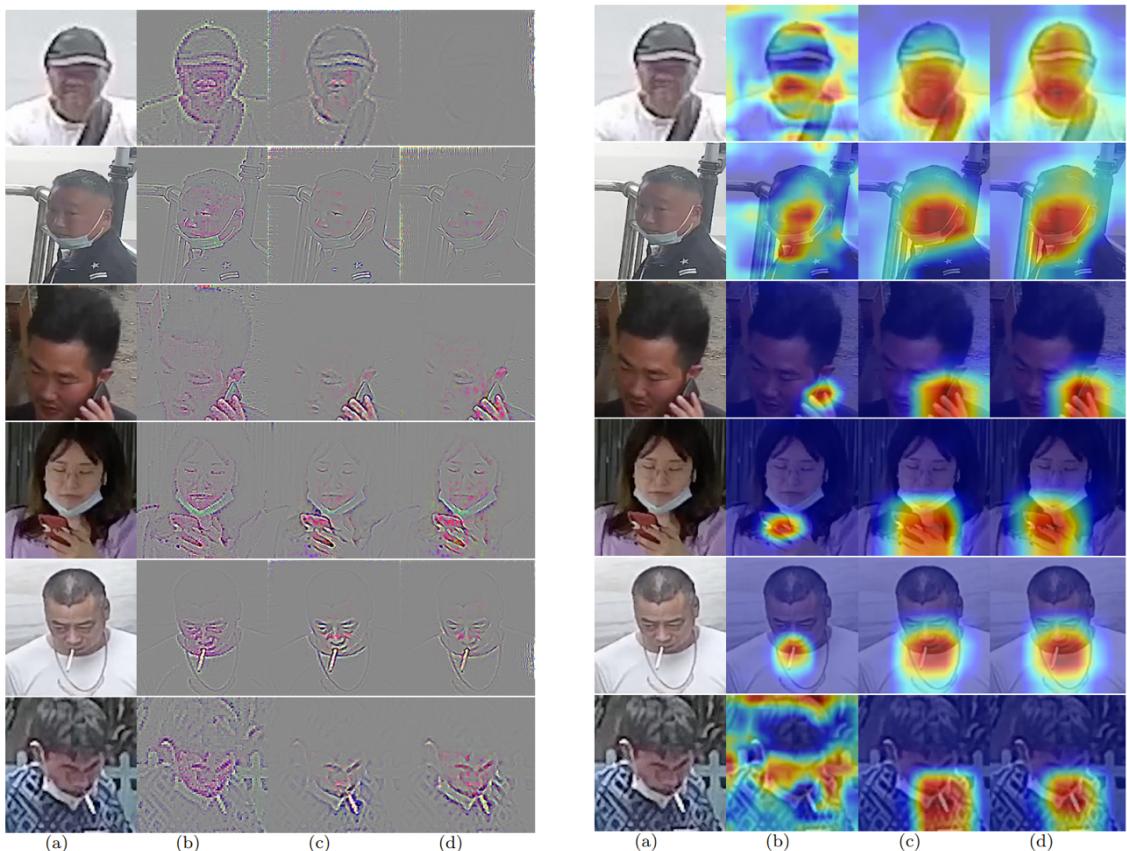
表 4. 是否使用数据增广在验证集上的 Top-1 准确度

表(4)中给出了在 SCResNet-50 模型基础上，是否使用数据增广对验证集 Top-1 准确度的影响。由结果可以看出，数据增广对模型性能的影响十分关键，直接带来了 4.53% 的准确度提升。对比未使用数据增广的 SCResNet-50 和表(2)中 SCVGG-17 的结果可以发现，在数据量比较少的情况下，模型的性能不占主导的地位，而只有在训练数据足够多的情况下，模型的优越性才会被表现出。

### 5.3. 网络可视化

虽然图像分类任务在深度学习领域中算是一个比较简单的任务，许多模型在 ImageNet 上的表现甚至也超越人类的判别水平，但对于网络中各模块的具体功能我们一直都不算清楚，模型可解释性欠佳。借助 GBP 和 Grad-CAM 技术，我们分别对 SCVGG-17、SCResNet-50 和 SCResNet-101 三个模型，在测试集 A 榜中进行了可视化探究，尝试解释网络的分类判别情况。

图(7)展示了三个模型分别在正常、吸烟和打电话类别上的 GBP 梯度可视化图像。很明显我们可以发现网络对图片的边缘比较敏感，例如人戴的帽子、人的眼睛，脸部等。通过对图(7)的第(b)、(c)、(d)三列可以发现，SCVGG-17 网络关注的区域比较广泛，而 SCResNet-50 和 SCResNet-101 网络关注的区域则更有针对性，对于吸烟和打电话的两个类别，可以很明显地发现嘴部的香烟以及人手持的手机部分的梯度值更大，说明被重点关注。也说明网络是通过判断是否有烟头或手持手机打电话这些特征来对图片进行判别的。



**图 7.** 导向反向传播在不同图片和不同网络下的可视化情况。列(a)是输入图片，列(b)是 SCVGG-17 输出结果，列(c)是 SCResNet-50 输出结果，列(d)是 SCResNet-101 的输出结果。

**图 8.** Grad-CAM 在不同图片和不同网络下的可视化情况。列(a)是输入图片，列(b)是 SCVGG-17 输出结果，列(c)是 SCResNet-50 输出结果，列(d)是 SCResNet-101 的输出结果。

借助 Grad-CAM 技术，我们可以得知图像中的各个部分对目标分类的“激活”情况。对于 SCVGG-17 网络，我们选取的是网络最后一个 ReLU 激活层进行可视化时；而对于 SCResNet-50 和 SCResNet-101 网络，我们选取的是第 4 个卷积块中的第 2 个 ReLU 激活层，即 layer4.2。

图(8)展示了三个模型分别在正常、吸烟和打电话类别上的 Grad-CAM 可视化图像。对比每一行的(b)、(c)、(d)列可以发现，SCVGG-17、SCResNet50 和 SCResNet-101 对特征的关注程度逐渐上升，且 SCVGG-17 关注的区域比较小，例在图(8)第 3 行和第 4 中，SCVGG-17 只关注了手机；而 SCResNet-50 和 SCResNet-101 还会关注持手机的手，一定程度上说明网络的泛化能力更好。在第 6 行中，SCVGG-17 的关注的区域有些错乱。整体上，说明网络会关注图片中是否含有烟头或是否存在手持手机，从而进行图片分类。

## 6. 总结与未来工作

在本文中，针对特殊场景下的吸烟打电话检测问题，我们视之为一个分类问题，基于 VGG 和 ResNet 的卷积神级网络结构，我们分别设计了适应本文任务的 SCVGG-17、SCResNet-50、SCResNet-101 和 SCResNet-151 网络。其中 SCResNet 在验证集上的 Top-1 准确率和 mAP 值均在 98% 以上，权衡模型参数量和模型分类准确率，可以认为 SCResNet-50 是一个综合性能较好的网络结构。

鉴于模型的解释性问题，本文借助导向反向传播 (GBP) 和 Grad-CAM 技术，分别对测试集的不同输入样本进行了可视化，对于吸烟和打电话类别的图片，结果表明网络会着重关注图像中的香烟和手持电话部分，这和人为进行图像分类的准则很相似。

虽然模型的性能总体而言较好，但还是有一些误分类的情况，我认为这和训练的数据量有一定的关系，本文训练集的数量为 5K 张图片，且图片主要来源于监控摄像头，图片的数量和场景来源相对于 ImageNet 这种大型图像库而言还是很少的，如想要部署到端设备上，后续可以考虑在更多的图像上进行训练，提高模型在各种场景下的分类泛化性能。

## 7. 课程总结

- 通过学习本次机器学习课程，我接触到了一些比较前沿的概念和技术，例如，CNN、RNN、LSTM、强化学习和宽度学习等，这也是当前计算机领域比较热门的方向。
- 通过本项目，我也对一些经典的分类网络展开了调研、学习及代码实现，虽然分类任务已经是一个比较成熟的技术，但其中的很多知识和论文都是其他领域的基础。此外，通过研读一些网络可视化相关的论文，我也对网络内部的一些机理有了更深刻的理解，例如梯度在网络中的作用及含义等。
- 最后，在论文的撰写部分，我也学会了用 LaTeX 来写论文，但课程论文最后要求改成 word 版本，希望以后以后论文格式这块不要卡太死。

## 参考文献

- [1] B. Chakraborty, B. Shaw, J. Aich, U. Bhattacharya, and S. K. Parui. Does deeper network lead to better accuracy: a case study on handwritten Devanagari characters. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 411–416. IEEE, 2018.
- [2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.
- [3] A. Dosovitskiy and T. Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4829–4837, 2016.
- [4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2010. In  
<http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html>, 2010.
- [5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [6] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.
- [8] D. G. Lowe. Distinctive image features from scaleinvariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

- [9] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.
- [10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- [11] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [12] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.
- [13] X.-Y. Xiao, R. Hu, S.-W. Zhang, and X.-F. Wang. Hog-based approach for leaf classification. In *International Conference on Intelligent Computing*, pages 149–155. Springer, 2010.