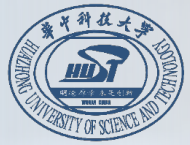


華中科技大學
Huazhong University of Science and Technology

神经网络可解释性

HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

明德厚学　求是创新

01 什么是可解释性

What's the Interpretability

什么是可解释性

- **Interpretability (of a DNN) is the ability to provide explanations in understandable terms to a human.**

  F Doshi-Velez & B Kim, 2017

  可解释性是指以可理解的方式向人类提供解释的能力

- **可解释性** (interpretability) 这个词主要是指解释具体的/已经训练好的网络。

  一个具体的神经网络无非是一个从输入x到输出y的非线性映射，**可解释性**就是想理解这个映射背后的「思路/逻辑/rationale」，而不是仅仅知道该结果是怎么经过一堆意义不明的数值（权值）计算出来的。

- **解释**（Explanations）：是指需要用某种语言来描述和注解
- **可解释的边界**（Explainable Boundary）：是指可解释性能够提供解释的程度
- **可理解的术语**（Understandable Terms）：是指构成解释的基本单元

为什么需要可解释性

- **高可靠性要求**

  神经网络在实际使用中经常被发现有一些意想不到的错误（更不用说对抗攻击），这对一些要求高可靠性的系统来说就很危险了（不信任）。可解释性可能有助于发现潜在的错误（比如发现模型逻辑和 domain knowledge 不相符），也可能可以帮助 debug，改进模型

- **伦理/法规要求**

  AI医疗：目前一般只作为辅助性的工具，是因为一个合格的医疗系统必须是透明的、可理解的、可解释的，可以获得医生和病人的信任；司法决策：面对纷繁复杂的事实类型，除了法律条文，还需要融入社会常识、人文因素等。因此，AI在司法决策的事后，必须要给出法律依据和推理过程。

- **作为其它科学研究的工具**

  科学研究是为了发现新知识。如果神经网络在某些科学问题上效果很好，那说明其可能发现了某种新「知识」，可解释性可以用来揭示它

卷积神经网络的可解释性分析

Interpretability Analysis of Convolutional Neural Networks

卷积神经网络的可解释性分析

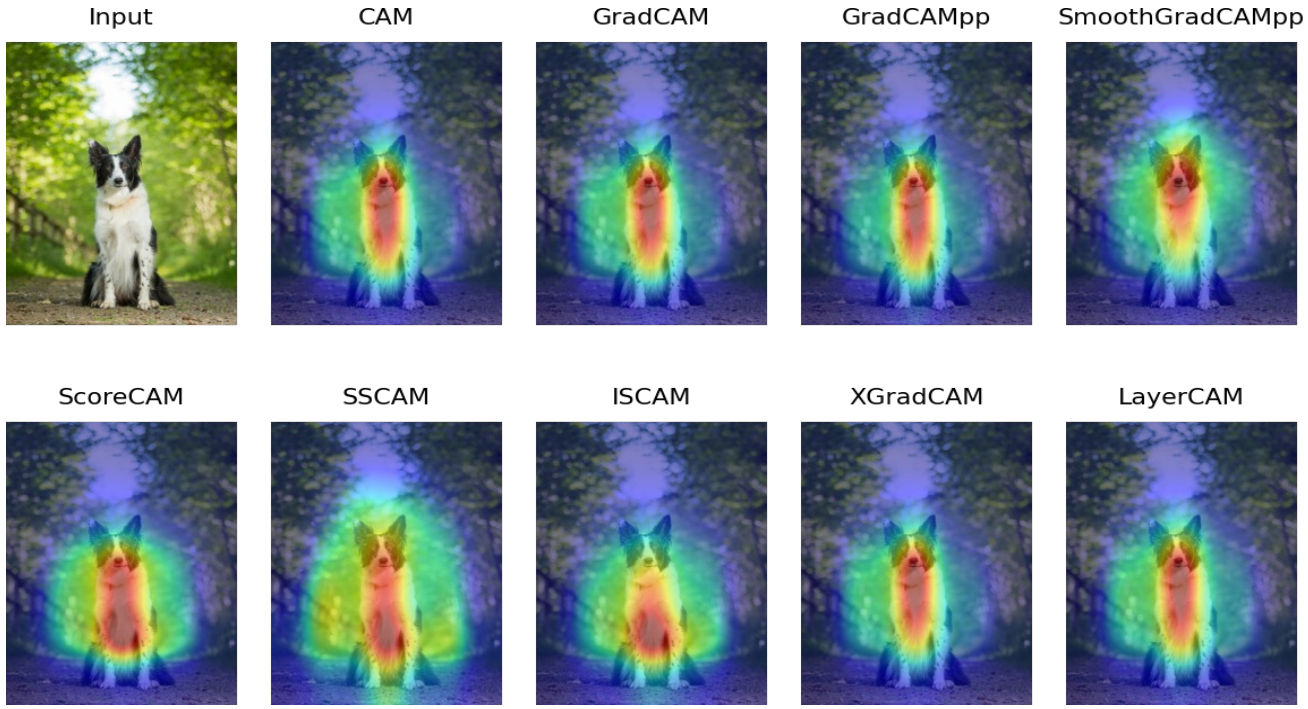
- 可视化卷积核、特征图
- 遮挡、缩放、平移、旋转
- 找到能使某个神经元激活的原图像素，或者小图
- **基于类激活热力图(CAM)的可视化**
- 语义编码降维可视化
- 由语义编码倒推输入的原图
- 生成满足某些要求的图像 (某类别预测概率最大)

基于类激活热力图(CAM)的可视化

Visualization based on Class Activation Map (CAM)

基于类激活热力图(CAM)的可视化

- 一直以来，深度神经网络的可解释性都被大家诟病，训练一个神经网络被调侃为“炼丹”。所得的模型也像一个“黑盒”一样，给它一个输入，然后得到结果，却不知道模型是如何得出结论的，究竟学习到了什么知识。如果能将其训练或者推理过程可视化，那么可以对其更加深入的理解。
- 而可视化**类激活热力图**能够更进一步地可视化神经网络在预测某一类别时，具体关注了图像的哪些像素。
- 右图是一个图像分类任务在几种CAM算法下的热力图，图中色块颜色越红，则说明神经网络对这块的关注度越大。



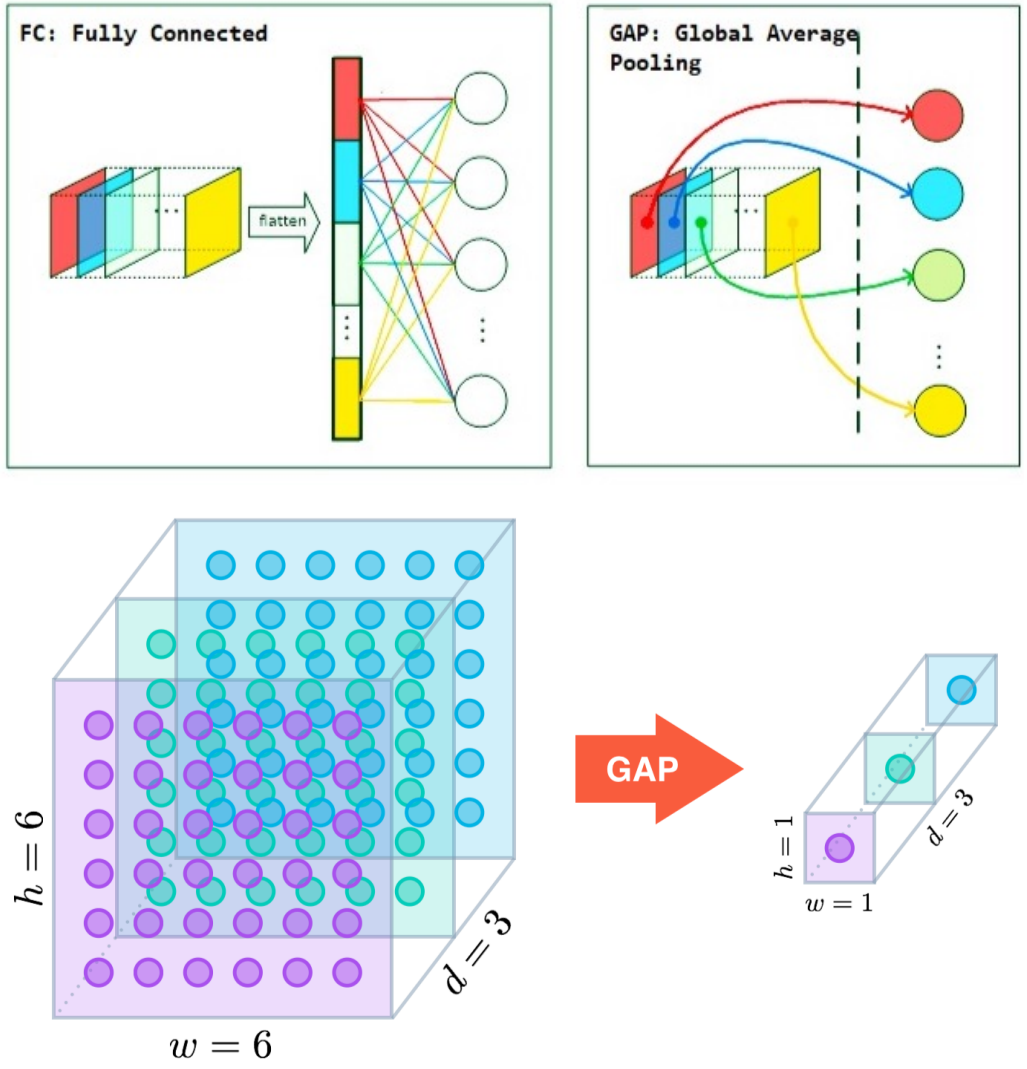
04 基于类激活热力图(CAM)的可视化

Visualization based on Class Activation Map (CAM)

全局平均池化（Global Average Pooling，GAP）

- 在常见的卷积神经网络中，全连接层之前的卷积层负责对图像进行特征提取，在获取特征后，**传统**的方法是接上**全连接层**之后再进行激活分类，而**GAP**的思路是使用**全局平均池化**来替代该全连接层(即使用池化层的方式来降维)，更重要的一点是保留了前面各个卷积层和池化层提取到的**空间信息**\语义信息，所以在实际应用中效果提升也较为明显。

- **全局平均池化**[1]使用一个标量来间接代表全卷积层的最后一层的一个Channel，具体的做法是对Channel取平均值。他可以取代全连接层，避免使用全连接层导致参数量爆炸，还可以减少参数量、防止过拟合。



[1] Lin M, Chen Q, Yan S. Network in network[J]. arXiv preprint arXiv:1312.4400, 2013.

04 基于类激活热力图(CAM)的可视化

Visualization based on Class Activation Map (CAM)

華中科技大學 Huazhong University of Science and Technology

类激活热力图(Class Activation Map, CAM)

- 全局平均池化层的简单修改与**类激活热力图**[2]（CAM）相结合，该技术允许经过分类训练的CNN同时进行分类和定位改类特定的图像区域例如用于刷牙的牙刷和用于砍树的链锯。从图中可以看到，神经网络对牙刷、链锯、人脸的部分注意力更集中，即神经网络通过这些特征对图像进行了分类。

刷牙

砍树

[2] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2921-2929.

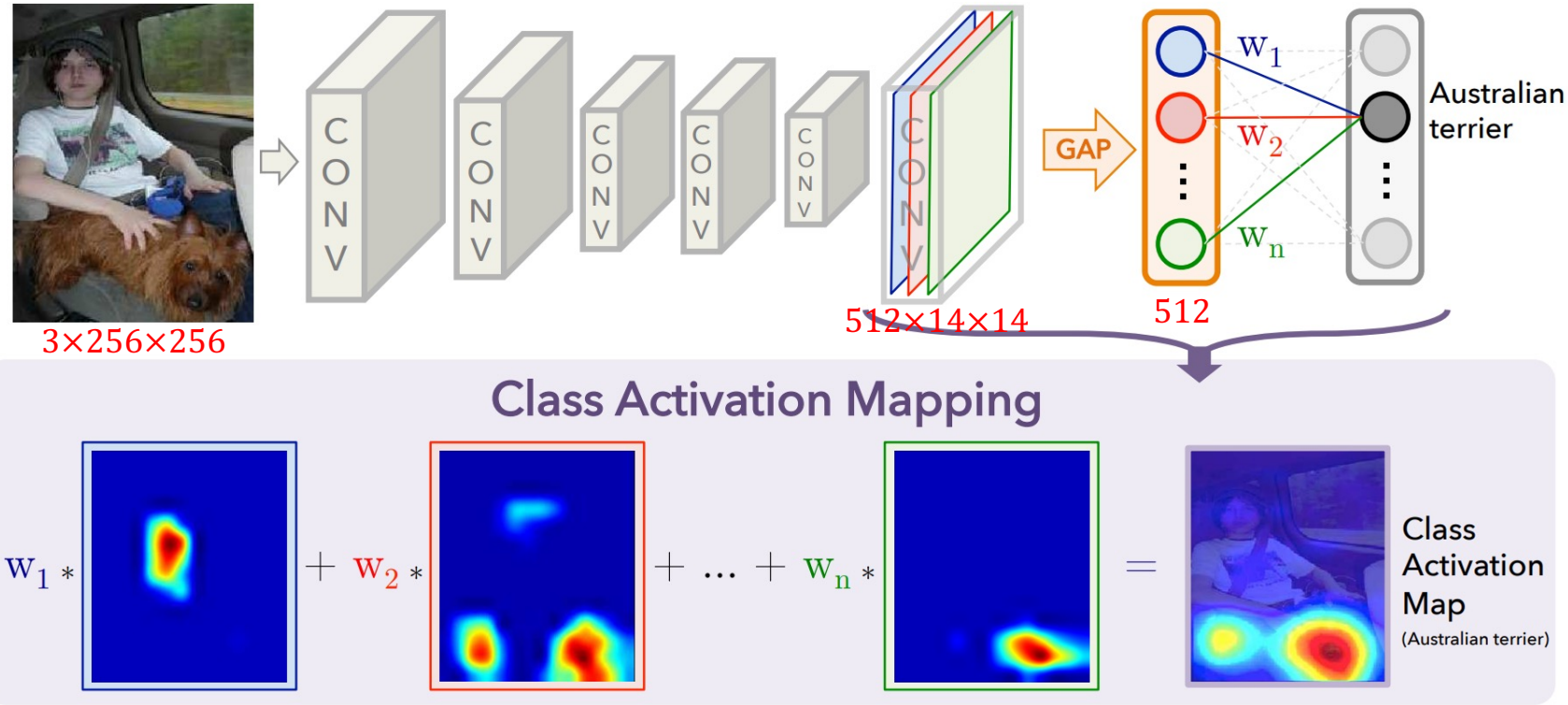
04 基于类激活热力图(CAM)的可视化

Visualization based on Class Activation Map (CAM)

華中科技大學 Huazhong University of Science and Technology

CAM算法原理

- 如下图是个分类网络，他的过程实际上就是一个新的CNN网络：从输入图像，到CNN网络，**最后一层的卷积层**（512个Channel）接到**GAP**。从GAP中可以得到512个平均值，即512个特征，最后的线性层实际上就是对得到的512个特征让神经网络去学习哪个的**权重**更大，最后得到最后的预测结果。
- 将最后训练的权重与最后一层卷积的图像进行加权求和后经过ReLU激活函数，最后再经过上采样（双线性插值）到与原始图像相同尺寸，既可以得到**类激活热力图**
- 如右图，神经网络分类结果为Australian terrier（澳大利亚梗，一种狗），则生成的热力图可以看到神经网络将大部分注意力集中在狗的身体和脸部。

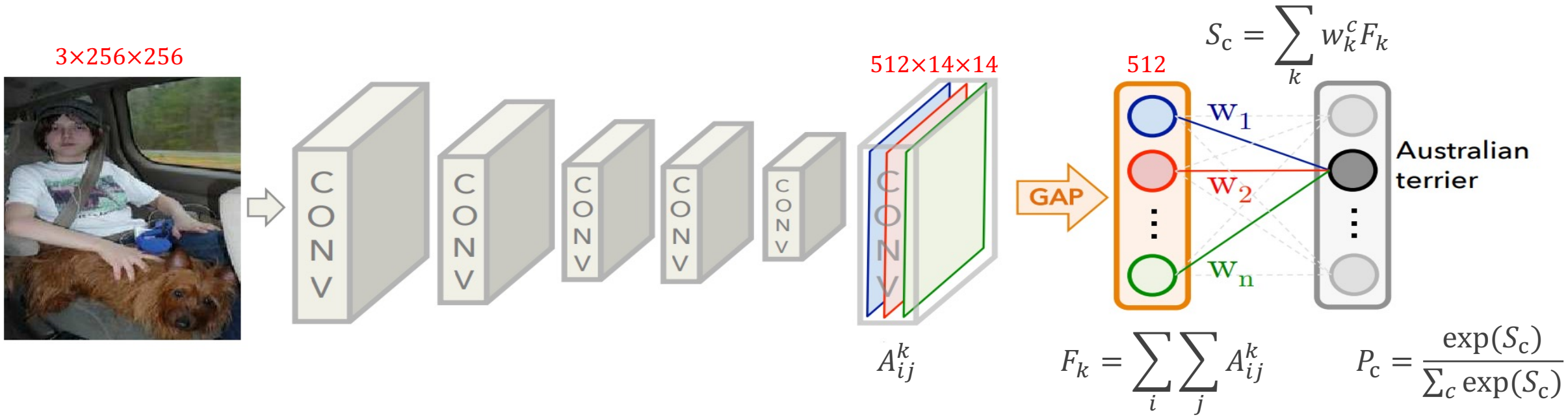


基于类激活热力图(CAM)的可视化

Visualization based on Class Activation Map (CAM)

CAM算法原理

- 如下图所示，关注最后一层卷积，其在第k个通道中，位置为(i,j)的值为A_{ij}^k，对于每一个通道k执行全局平均池化（GAP），得到$F_k = \sum_i \sum_j A_{ij}^k$。最后对于分类$c$，输入到$softmax$层的数据$S_c = \sum_k w_k^c F_k$，其中$w_k^c$表示GAP结果$F_k$到数据$S_c$之间全连接层的权重。最后$softmax$层的输出$P_c = \frac{\exp(S_c)}{\sum_c \exp(S_c)}$。
- 因此可以看出，权重w_k^c可以反映每一个F_k对于分类c的重要程度。



04 基于类激活热力图(CAM)的可视化

Visualization based on Class Activation Map (CAM)

華中科技大學 Huazhong University of Science and Technology

CAM算法原理

- $Softmax$的$bias$被定义为0，因为它对分类几乎没有影响。如下公式，定义M_c，则$ReLU(M_c)$为分类c的类激活热力图，它直接反映了将图像分类为类别c的依据，它反映了图像在空间(i,j)中的重要性，导致了图像被分类为c。此方法得到的热力图和原图尺寸不一致，可以考虑使用双线性插值或其他方法缩放到原图尺寸。

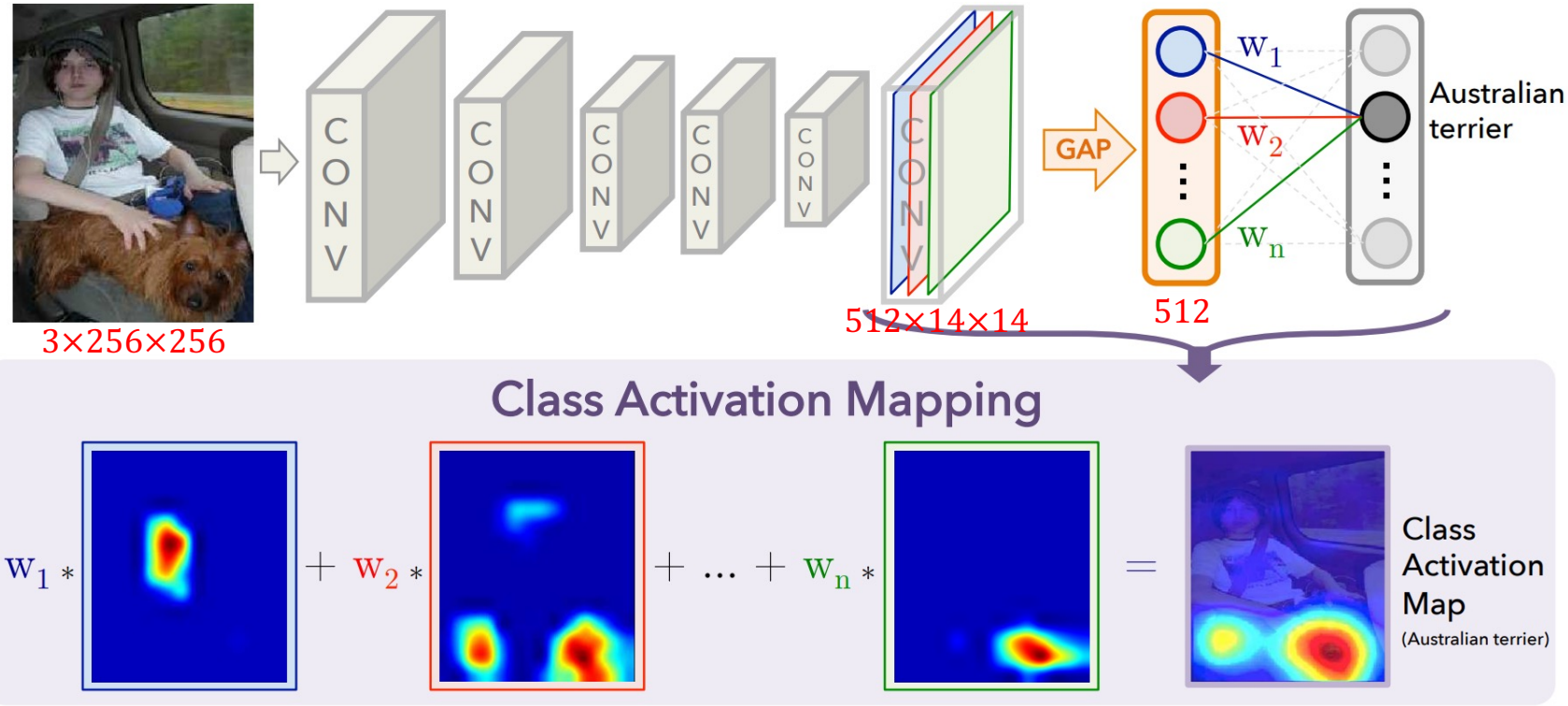
$$S_c = \sum_k w_k^c \sum_i \sum_j A_{ij}^k$$

$$= \sum_i \sum_j \sum_k w_k^c A_{ij}^k$$

$$M_c(i,j) = \sum_k w_k^c A_{ij}^k$$

$$S_c = \sum_i \sum_j M_c(i,j)$$

$$L_{CAM}^c = ReLU(\sum_k w_k^c A_{ij}^k)$$

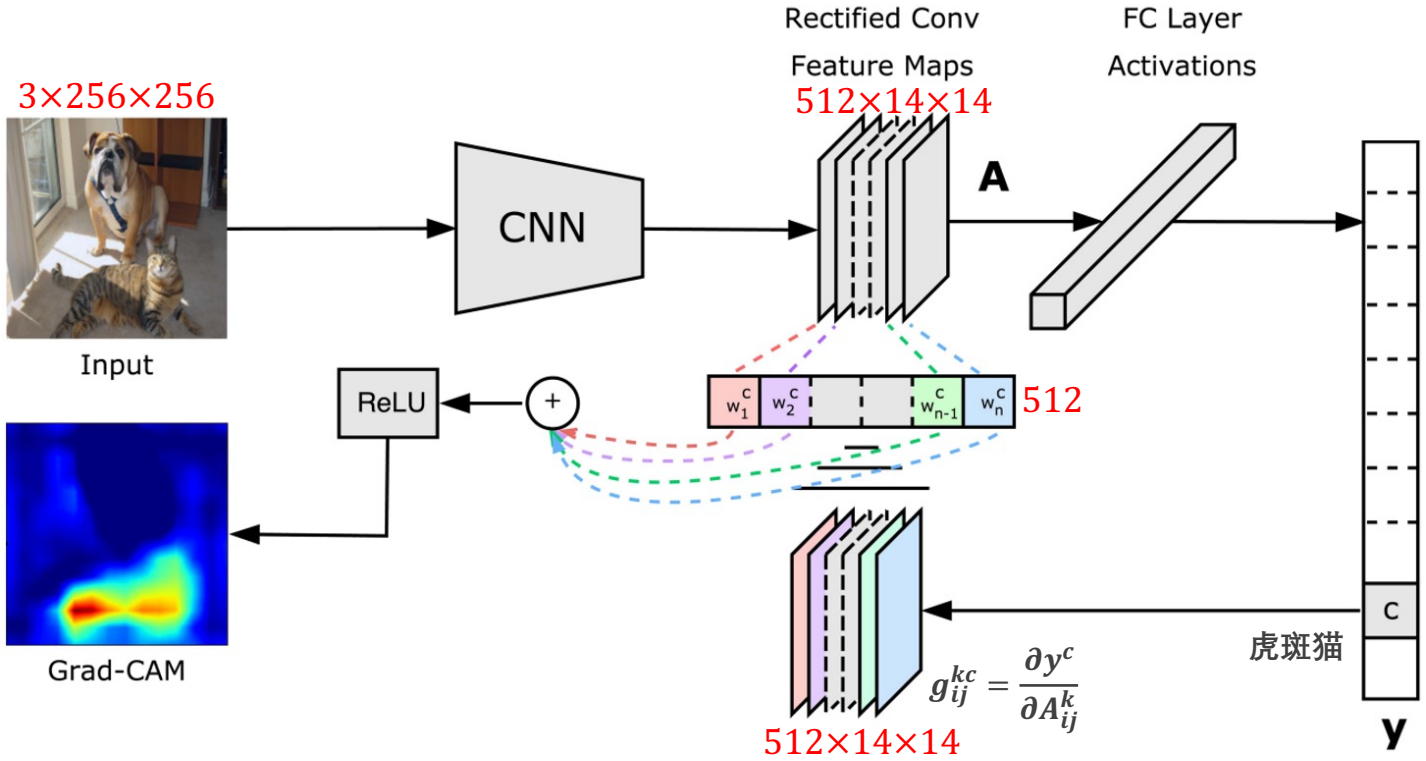


基于类激活热力图(CAM)的可视化

Visualization based on Class Activation Map (CAM)

Grad-CAM（Gradient Weighted Class Activation Map）

- CAM得到的效果已经很不错了，但其要求模型必须有GAP，否则需要修改网络结构并对模型进行重新训练，这样就导致其应用起来很不方便。此外，它还有一些缺点：只能分析最后一层卷积层输出，无法分析中间层、仅限图像分类任务
- **Grad-CAM**[3]不要求模型必须有GAP层，它对输出的分数进行求导获得最后一层卷积层的梯度，对梯度进行GAP，得到权重，与最后一层卷积的特征图进行加权求和后经过ReLU激活函数，最后再经过上采样（双线性插值）到与原始图像相同尺寸，既可以得到**类激活热力图**
- **Grad-CAM**不修改网络结构，不需要重新训练，只要模型输出值可以求导，就能应用于各种模型。也可以应用于中间的卷积层，但卷积层的层数越浅热力图效果较差。



[3] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE international conference on computer vision. 2017: 618-626.

基于类激活热力图(CAM)的可视化

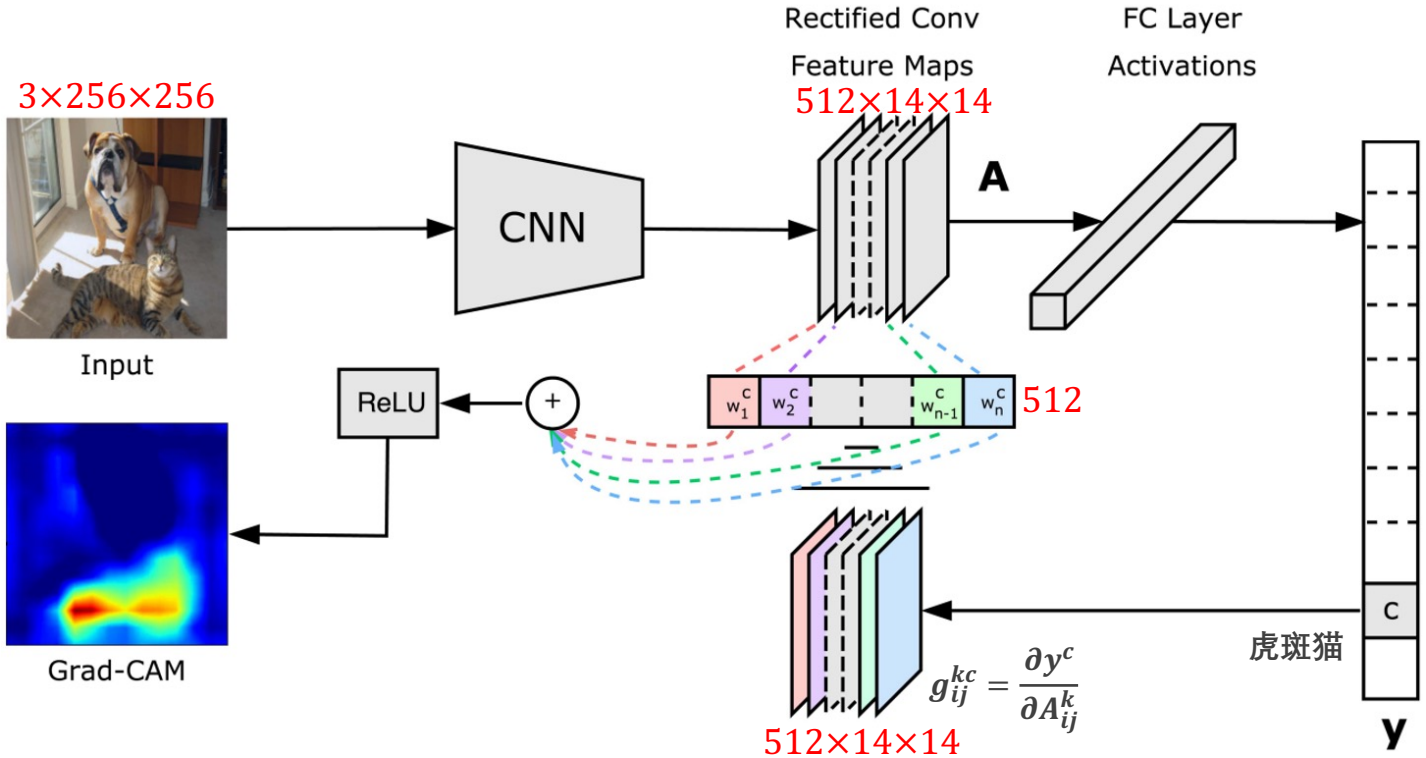
Visualization based on Class Activation Map (CAM)

Grad-CAM（Gradient Weighted Class Activation Map）

- 在分类模型中，最后一层卷积为矩阵A，通道k中位置为(i,j)的位置被定义为A_{ij}^{k}。最后一层卷积经过全连接层后得到各分类的分数（送入$softmax$前），其中分类c的分数为y^{c}。
- Grad-CAM[3]不要求模型必须有GAP层，求得分数y^{c}对最后一层卷积A_{ij}^{k}每一个元素的偏导数$g_{ij}^{kc}=\frac{\partial y^{c}}{\partial A_{ij}^{k}}$。经过全局平均池化得到$w_{k}^{c}=\frac{1}{N}\sum_{i}\sum_{j}\frac{\partial y^{c}}{\partial A_{ij}^{k}}$，其中$N$为$A_{ij}^{k}$的像素数。类似CAM，可以得到热力图：

$$L_{Grad-CAM}^{c}=ReLU(\sum_{k}w_{k}^{c}A_{\mathrm{ij}}^{k})$$

- 其中y^{c}可以替换为网络中任何一个可微分的值，因此Grad-CAM也可以分析除了最后一层卷积以外的其它层



基于类激活热力图(CAM)的可视化

Visualization based on Class Activation Map (CAM)

LayerCAM

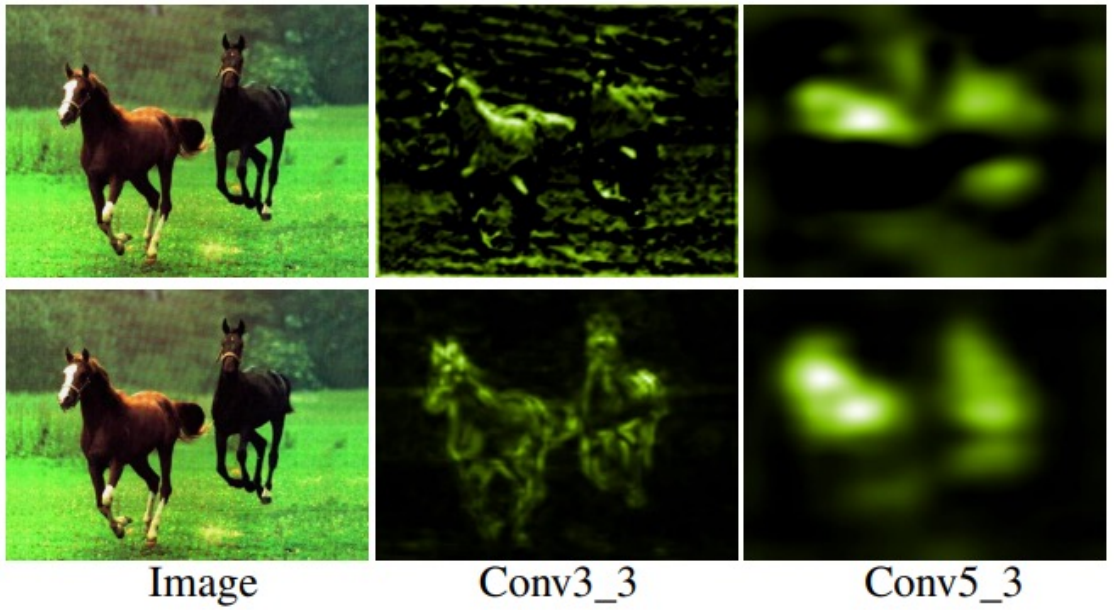
- 然而Grad-CAM还有很多缺点，**LayerCAM**[4]的提出解决了其中的一部分缺点：(1)深层生成的粗粒度热力图和浅层生成的细粒度热力图都不够精确；(2)GradCAM可以分析中间层，但发现在浅层中，分析效果很差。
- 如下图所示，其中第一行为GradCAM生成的热力图，在浅层中难以分析到关注的具体区域，第二行为LayerCAM生的热力图，可以看到即使在浅层中也能看到很明显的关注区域。
- 对于任意一层CNN，LayerCAM针对分类结果y^c对其每一个Channel进行反向传播得到梯度，并计算权重：

$$w_{ij}^{kc} = \mathrm{ReLU}(g_{ij}^{kc})$$

- 得到热力图：

$$L_{LayerCAM}^{c} = ReLU(\sum_{k} w_{ij}^{kc} A_{ij}^{k})$$

- Grad-CAM对特征图的梯度进行全局平均池化得到一个标量的权重，而LayerCAM是对特征图的梯度进行元素级的乘法，从而生成更精细的类别激活图



[4] Jiang P T, Zhang C B, Hou Q, et al. Layercam: Exploring hierarchical class activation maps for localization[J]. IEEE Transactions on Image Processing, 2021, 30: 5875-5888.

華中科技大學

Huazhong University of Science and Technology

THANKS

明德厚学　求是创新