可解释深度学习技术在医疗图像诊断中的应用

刘磊

背景

现有的深度学习模型中的主要参数和结构都不能直接解释模型。

深度学习方法的黑盒特性是阻碍医疗应用的主要原因。

医学模型的应用需要大量的人工监督。

定义

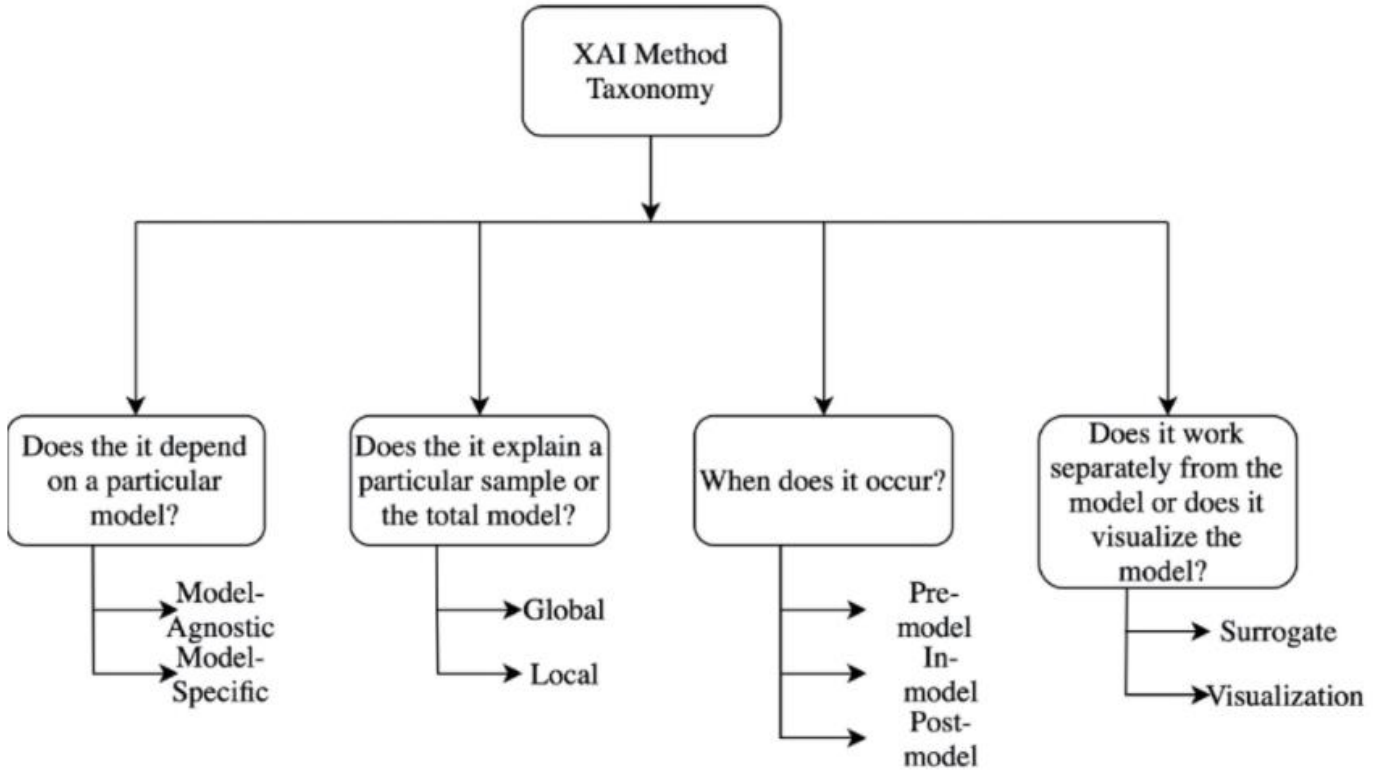
深度学习方法的可解释性是指能够理解深度学习模型内部机制以及能够理解深度学习模型的结果。

Interpretability: 将一个抽象概念(如输出类别)映射到一个域示例(Domain Example)。

Explainability: 指能够生成一组域特征(Domain Features)，如图像的像素，这些特征有助于模型的输出决策

医疗图像可解释性：一个医疗诊断系统必须是透明的、可理解的、可解释的以获得医生、监管者和病人的信任。理想情况下，它应该能够向所有相关方解释做出某个决定的完整逻辑。

可解释性方法分类



1. 模型特定方法 vs 模型无关方法（Model Specific vs Model Agnostic）
2. 全局方法 vs 局部方法（Global Methods vs Local Methods）
3. 模型前 vs 模型中 vs 模型后方法（Pre-model vs in-model vs post-model）
4. 替代方法 vs 可视化方法（Surrogate Methods vs Visualization Methods）

模型特定方法 vs 模型无关方法

模型特定的方法基于单个模型的参数进行解释。
模型无关方法不局限于特定的模型体系结构。这些方法不能直接访问内部模型权重或结构参数，主要适用于事后分析。

全局方法 vs 局部方法

局部可解释性方法主要聚焦于模型的单个输出结果，一般通过设计能够解释特定预测或输出结果的原因的方法来实现。
全局方法通过利用关于模型、训练和相关数据的整体知识聚焦于模型本身，试图从总体上解释模型的行为。特征重要性是全局方法的一个很好的例子，它试图找出在所有不同的特征中对模型性能有更好影响的特征。

模型前 vs 模型中 vs 模型后方法

模型前：独立于深度学习模型结构，如PCA 和 t-SNE
模型中：集成在深度学习模型中的方法。
模型后：主要关注模型在训练过程中学习到了什么。

替代方法 vs 可视化方法

替代方法由不同的模型组成一个整体，用于分析其他黑盒模型。通过比较黑盒模型和替代模型来解释替代模型的决策，从而辅助理解黑盒模型。
可视化方法：CAM

可解释深度学习模型在医疗图像分析中的应用分类

具体到医疗图像分析领域，引入可解释性方法的可解释深度学习模型主要有两类：属性方法和非属性方法（non-attribution based）。两类方法的主要区别在于是否已经确定了输入特征对目标神经元的联系。

属性方法的目标是直接确认输入特征对于深度学习网络中目标神经元的贡献程度。

---- Attribution Maps (heatmaps)

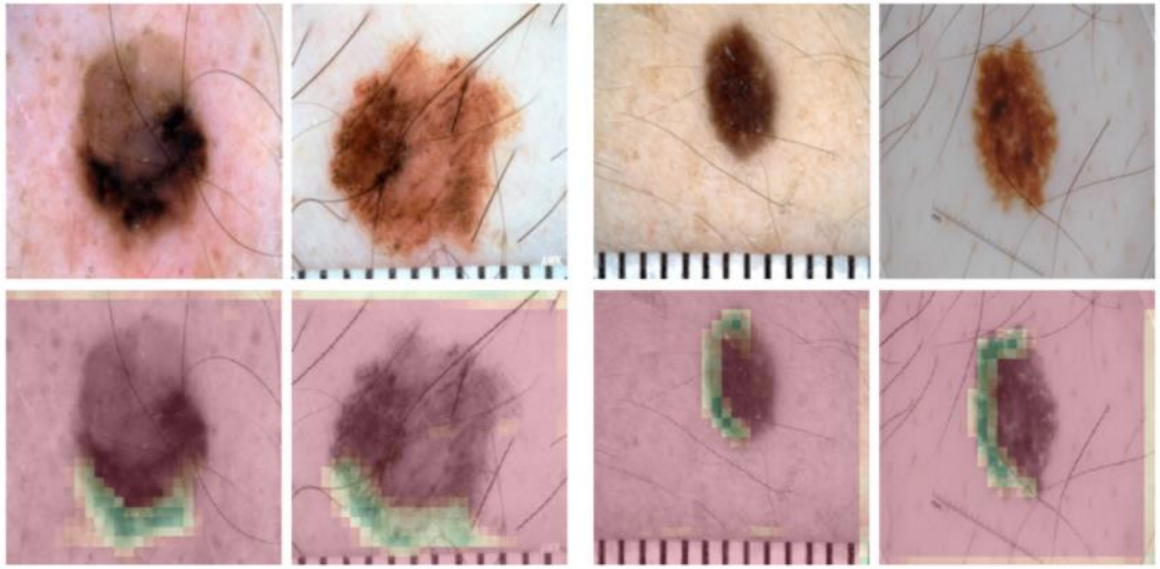
----通过扰动分析输入特征的改变对深度学习模型输出的影响。

非属性方法则是针对给定的专门问题开发并验证一种可解释性方法，例如生成专门的注意力、知识或解释性去辅助实现专门问题的可解释深度学习。

---- 注意力机制

Visualizing Convolutional Neural Networks to Improve Decision Support for Skin Lesion Classification

医生的诊断经验: 边界不规则的皮肤病变边界可能表明存在恶性病变。

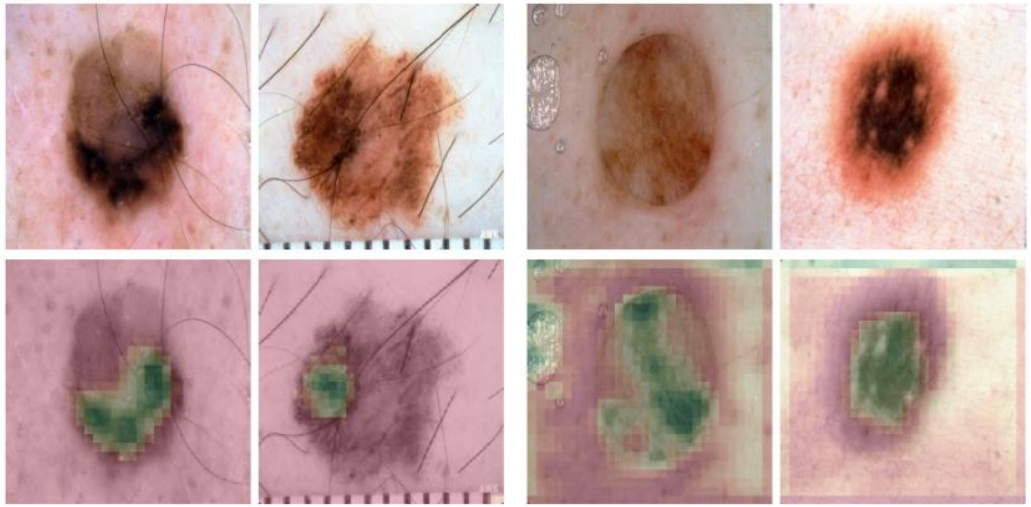


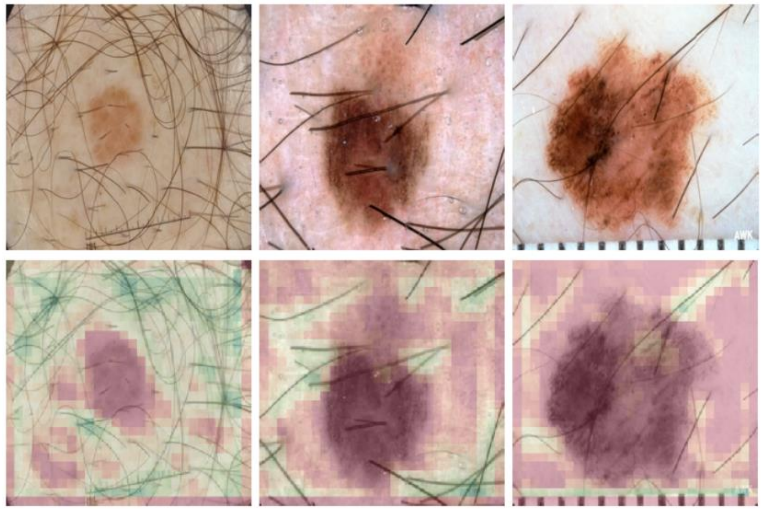
特征图在皮肤病变的边界上都有很高的激活率，但都处于边界的不同部位。

Visualizing Convolutional Neural Networks to Improve Decision Support for Skin Lesion Classification

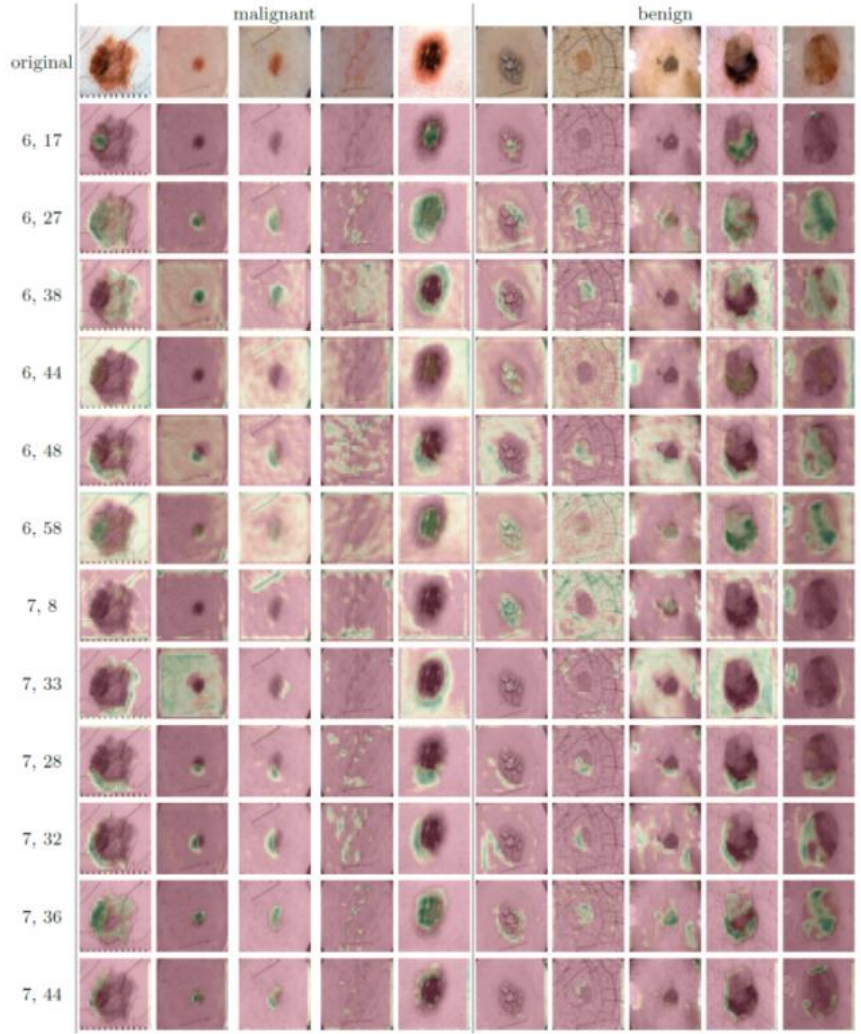
颜色均匀的病变通常是良性的，而严重的颜色不规则可能是恶性病变的征兆。

从皮肤科医生的角度考虑，头发对于最终的诊断没有影响。





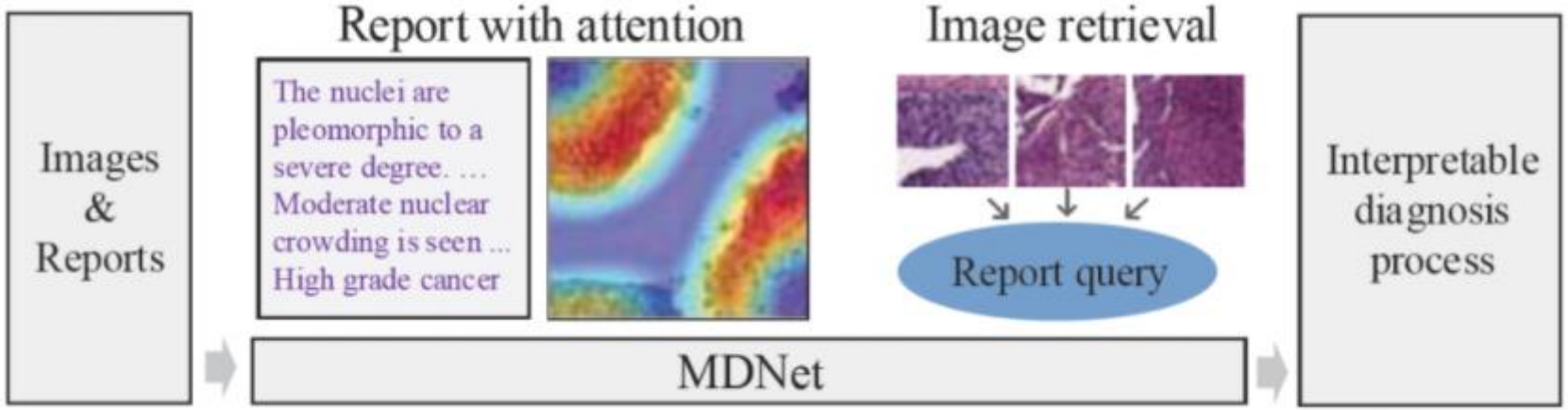
Visualizing Convolutional Neural Networks to Improve Decision Support for Skin Lesion Classification



尽管本文对 CNN 学习到的特征给出了一些分析和评论，但并不能解释 CNN 检测到的特征与其输出之间的任何因果关系。此外，通过特征图，并没有发现任何能精确突出皮肤科医生扫描过程中重点关注的其他结构，如球状体、圆点、血管结构等。作者认为，为了使 CNN 能够成为皮肤科医生更好的决策支持工具，还需要在这一领域进行更多的研究。

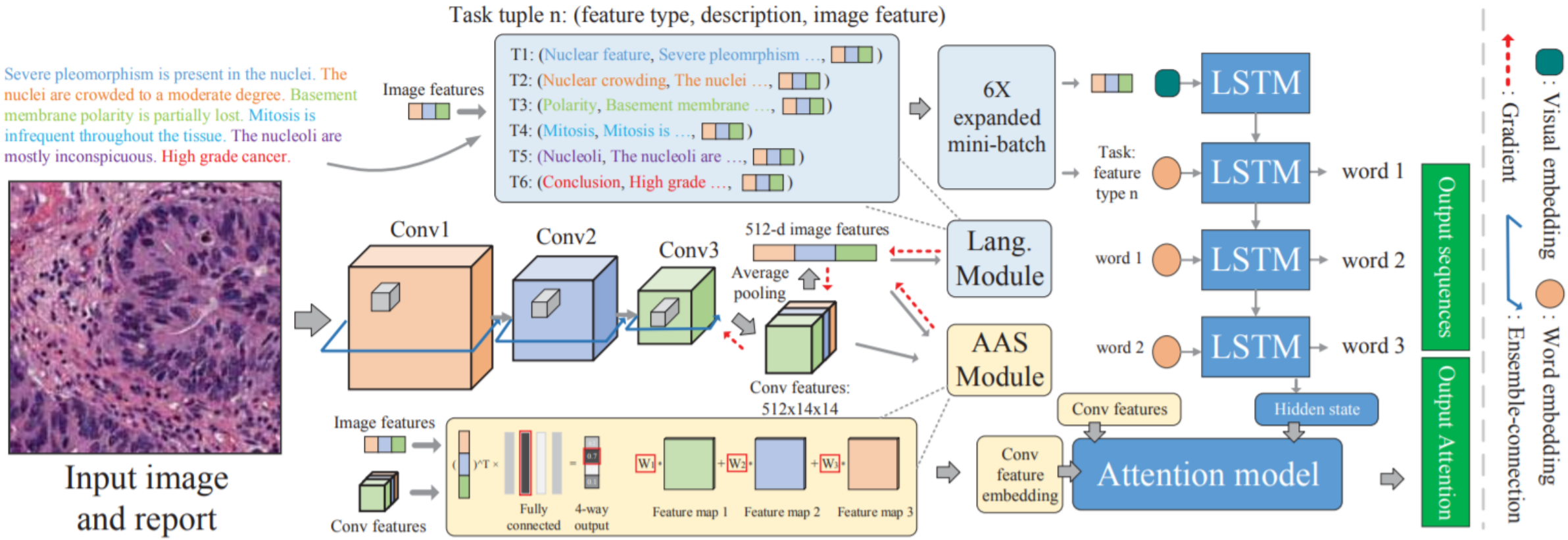
MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network

MDNet读取图像，生成诊断报告，通过症状描述检索图像，并将网络注意力可视化，通过建立医学图像与诊断报告之间的直接多模态映射为网络诊断过程提供依据。



MDNet：图像模块(CNN)+语言模块(LSTM+attention)

MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network



MDNet：图像模块(CNN)+语言模块(LSTM+attention)

MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network

$$y_L = y_l + \sum_{m=l}^{L-1} \mathcal{F}_m(y_m), \qquad \frac{\partial \mathcal{L}}{\partial y_l} = \frac{\partial \mathcal{L}}{\partial y_L}(1 + \frac{\partial}{\partial y_l} \sum_{m=l}^{L-1} \mathcal{F}_m(y_m)),$$

Fm = Conv + BN +ReLU (残差块)

连接卷积层的分类模块包括全局平均池化层和全连接层

$$p^c = \sum_k w_k^c \cdot \sum_{i,j} y_L^{(k)}(i,j),$$

y_L^k为最后一个残差块的第k个特征图，c为对应的类别。

MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network

$$p^c = \sum_k w_k^c \cdot \sum_{i,j} y_L^{(k)}(i,j), \quad \text{单一加权函数}$$

$$p^c = \sum_{i,j} \left(\boldsymbol{w}_1^c \cdot y_1 + \sum_{m=1}^{L-1} \boldsymbol{w}_{m+1}^c \cdot \mathcal{F}_m \right). \quad \text{解耦输出}$$

$$y_{l+1} = \mathcal{F}_l(y_l) \otimes y_l \quad \otimes\text{代表连接操作}$$

MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network

语言模型中，我们常常采用LSTM来进行输出诊断报告，这里输出的时候采用的是输出的单词在单词表中的概率。最大化句子的联合概率来建模诊断报告:

$$\log p(\boldsymbol{x}_{0:T}|I;\theta_L) = \sum_{t=0}^{T} \log p(\boldsymbol{x}_t|I, \boldsymbol{x}_{0:t-1};\theta_L),$$

x为一个句子的每个单词，为one-hot向量。Theta为LSTM的参数。

LSTM的输入: 根据图像特征和权值进行计算得到的z和语义特征h

下个单词的概率为:

$$\boldsymbol{h}_t = LSTM(E\boldsymbol{x}_{t-1}, \boldsymbol{h}_{t-1}, \boldsymbol{z}_t),$$
$$p(\boldsymbol{x}_t|I, \boldsymbol{x}_{0:t-1};\theta_L) \propto \exp(G_h\boldsymbol{h}_t),$$

E是词向量矩阵，G是输出映射

MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network

原来的注意力机制难以训练，因为本文提出一个锐化模块

$$\boldsymbol{a}_t = softmax(W_{att}\ \tanh(W_h\,\boldsymbol{h}_{t-1} + \boldsymbol{c})),$$
$$\boldsymbol{c} = (\boldsymbol{w}^c)^T C(I),$$
$$\boldsymbol{z}_t = \boldsymbol{a}_t\, C(I)^T,$$

W为学习到的词向量矩阵，C是卷积特征图

MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network

θ_D ： image model D

θ_L ： laguage model L

θ_M ： AAS model M （auxiliary attention sharpening）

损失函数

$$\max_{\theta_L,\theta_D,\theta_M} \mathcal{L}_M(l_c, M(D(I;\theta_D);\theta_M)) + \mathcal{L}_L(l_s, L(D(I;\theta_D);\theta_L)),$$

$$\theta_D \leftarrow \theta_D - \lambda \cdot \left((1-\beta) \cdot \frac{\partial \mathcal{L}_M}{\partial \theta_D} + \beta \cdot \eta \frac{\partial \mathcal{L}_L}{\partial \theta_D}\right)$$