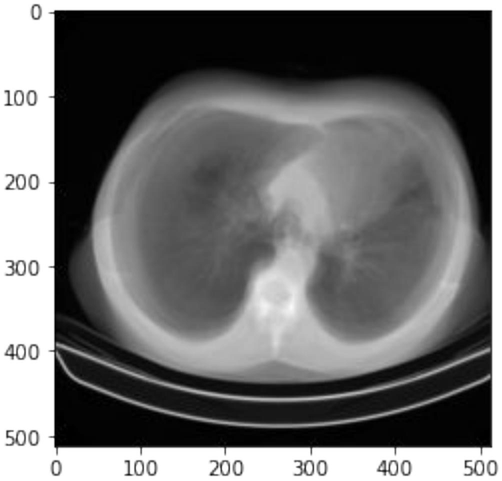
DIP项目答辩

李嘉杰 叶莱 贾小玉 周楷彬

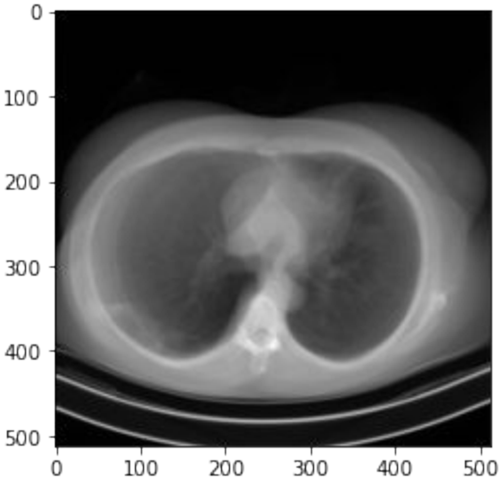
一、探索性数据分析(EDA)

EDA

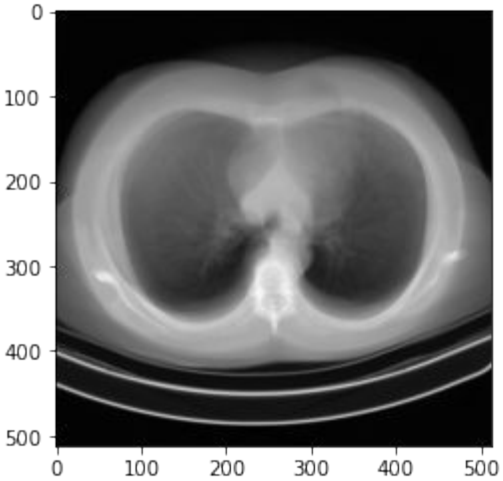
mean



normal



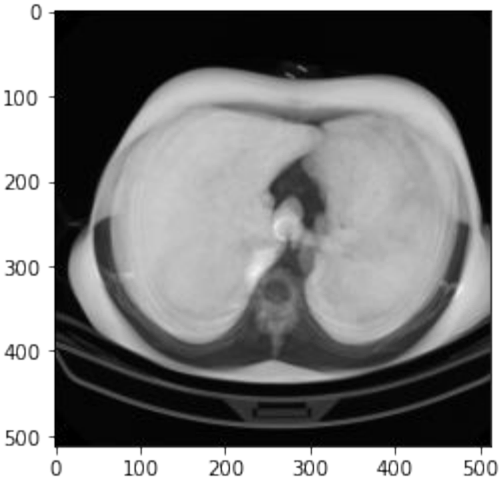
cap



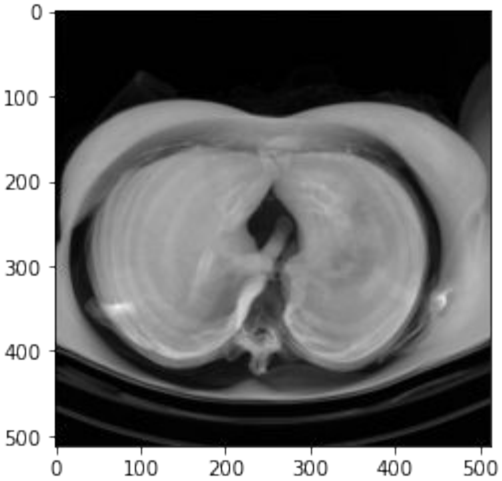
covid-19

EDA

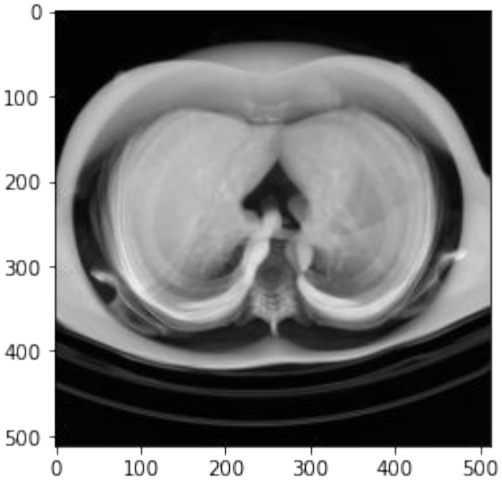
standard deviation



normal



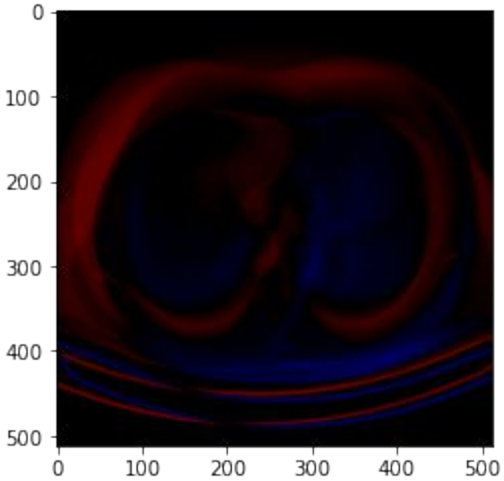
cap



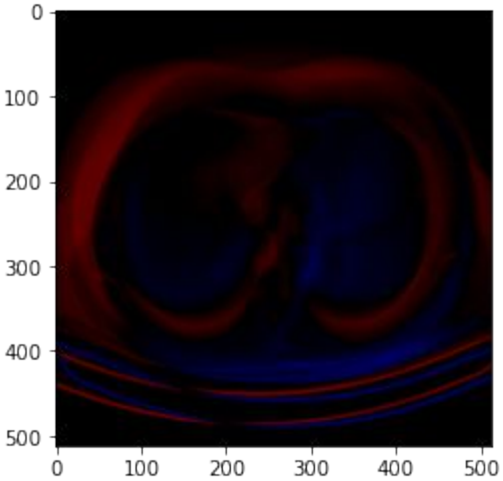
covid-19

EDA

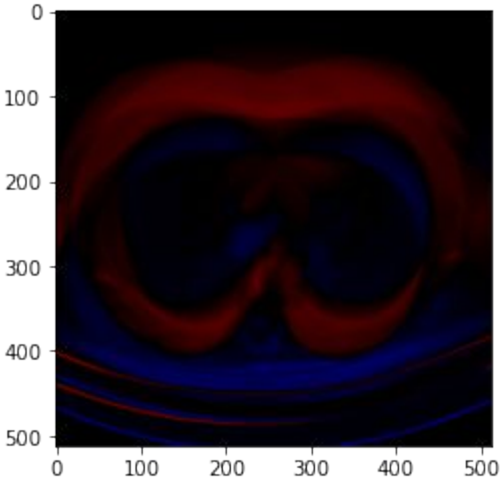
difference



cap & normal



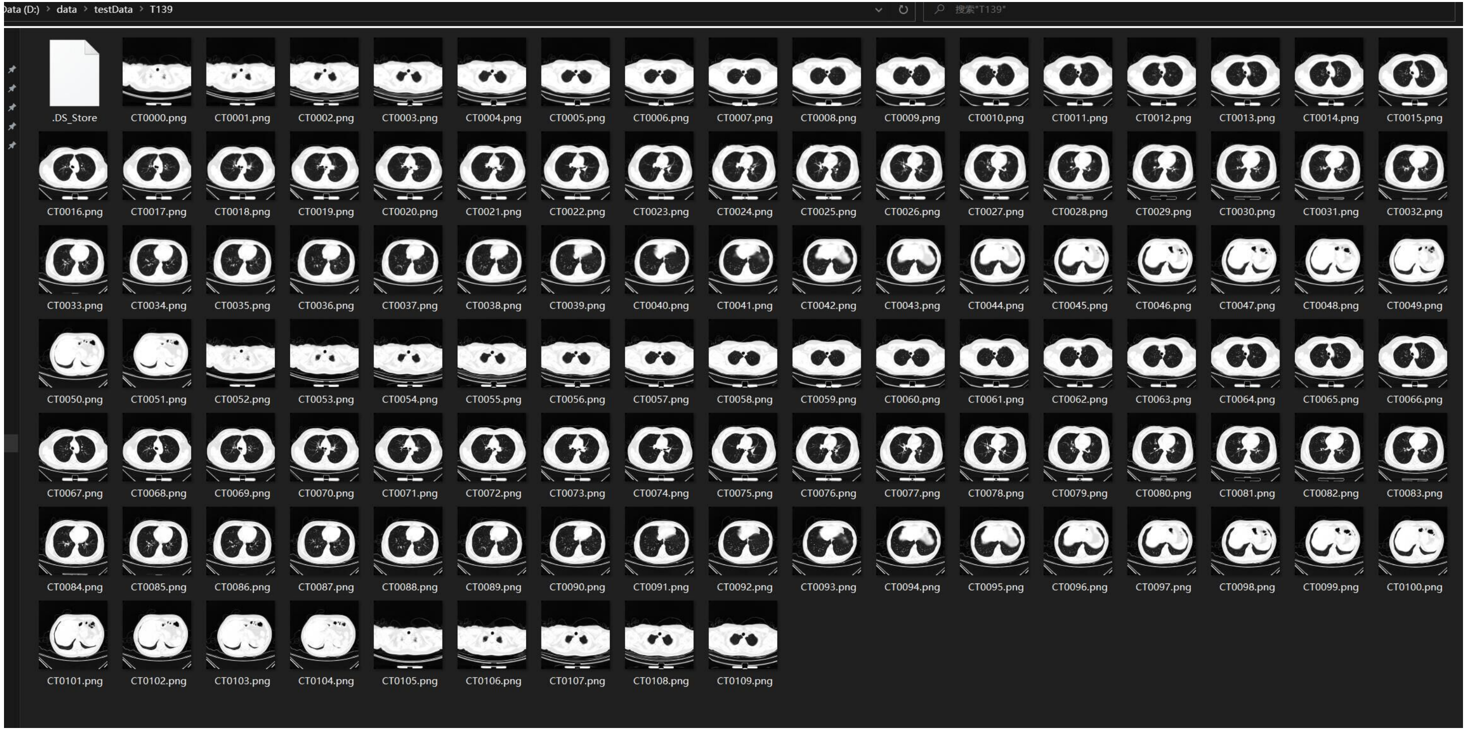
covid-19 & normal



covid-19 & cap

二、数据清洗和预处理

数据清洗-数据集混乱


ata (D:) > data > testData > T139
搜索"T139"
.DS_Store
CT0000.png
CT0001.png
CT0002.png
CT0003.png
CT0004.png
CT0005.png
CT0006.png
CT0007.png
CT0008.png
CT0009.png
CT0010.png
CT0011.png
CT0012.png
CT0013.png
CT0014.png
CT0015.png
CT0016.png
CT0017.png
CT0018.png
CT0019.png
CT0020.png
CT0021.png
CT0022.png
CT0023.png
CT0024.png
CT0025.png
CT0026.png
CT0027.png
CT0028.png
CT0029.png
CT0030.png
CT0031.png
CT0032.png
CT0033.png
CT0034.png
CT0035.png
CT0036.png
CT0037.png
CT0038.png
CT0039.png
CT0040.png
CT0041.png
CT0042.png
CT0043.png
CT0044.png
CT0045.png
CT0046.png
CT0047.png
CT0048.png
CT0049.png
CT0050.png
CT0051.png
CT0052.png
CT0053.png
CT0054.png
CT0055.png
CT0056.png
CT0057.png
CT0058.png
CT0059.png
CT0060.png
CT0061.png
CT0062.png
CT0063.png
CT0064.png
CT0065.png
CT0066.png
CT0067.png
CT0068.png
CT0069.png
CT0070.png
CT0071.png
CT0072.png
CT0073.png
CT0074.png
CT0075.png
CT0076.png
CT0077.png
CT0078.png
CT0079.png
CT0080.png
CT0081.png
CT0082.png
CT0083.png
CT0084.png
CT0085.png
CT0086.png
CT0087.png
CT0088.png
CT0089.png
CT0090.png
CT0091.png
CT0092.png
CT0093.png
CT0094.png
CT0095.png
CT0096.png
CT0097.png
CT0098.png
CT0099.png
CT0100.png
CT0101.png
CT0102.png
CT0103.png
CT0104.png
CT0105.png
CT0106.png
CT0107.png
CT0108.png
CT0109.png

数据清洗-数据集混乱


CT0066.png
CT0067.png
CT0068.png
CT0069.png
CT0070.png
CT0071.png
CT0072.png
CT0073.png
CT0074.png
CT0075.png
CT0076.png
CT0077.png
CT0078.png
CT0079.png
CT0080.png
CT0081.png
CT0082.png
CT0083.png
CT0084.png
CT0085.png
CT0086.png
CT0087.png
CT0088.png
CT0089.png
CT0090.png
CT0091.png
CT0092.png
CT0093.png
CT0094.png
CT0095.png
CT0096.png
CT0097.png
CT0098.png
CT0099.png
CT0100.png
CT0101.png
CT0102.png
CT0103.png
CT0104.png
CT0105.png

数据清洗-图像相似度

SSIM

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

PSNR

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) = 20 \cdot \log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right)$$

ImageHash: perception hashing, average hashing, difference hashing, wavelet hashing

数据清洗-图像相似度

image hashing:用于图像相似度计算。

·缩小尺寸。

·简化色彩。

·计算平均值。

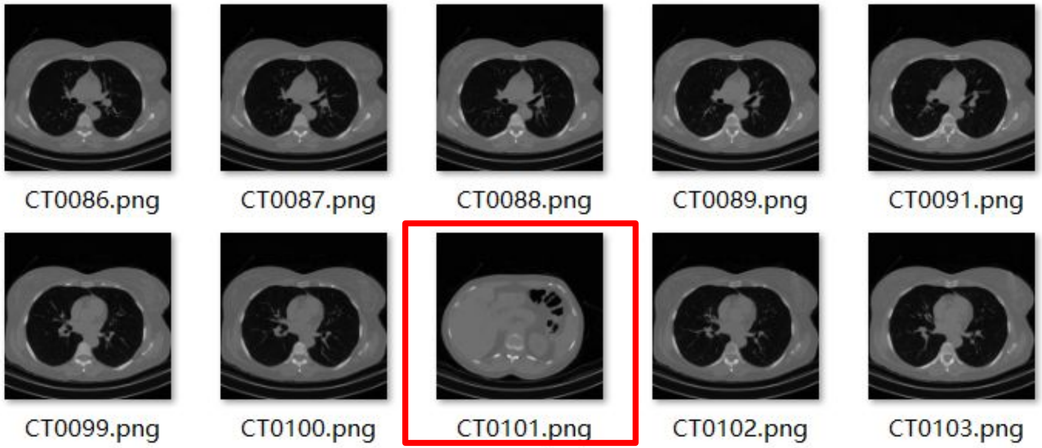
·比较像素的灰度。

·计算哈希值。

数据清洗-图像相似度

if **distance**(left,this)>5 and **distance**(this,right)>5:

remove(this)



去除背景-形态学操作

肺和气道的粗分割

1.强度阈值和形态运算

2.估计肺叶视野

3.选择最大的连接组件并保存图像

4.显示细分结果

//形态学操作

```
M = np.zeros(I.shape)
M[I > params['lungMinValue']] = 1
M[I > params['lungMaxValue']] = 0

struct_s = ndimage.generate_binary_structure(3, 1)
struct_m = ndimage.iterate_structure(struct_s, 2)
struct_l = ndimage.iterate_structure(struct_s, 3)
M = ndimage.binary_closing(M, structure=struct_s, iterations = 1)
M = ndimage.binary_opening(M, structure=struct_m, iterations = 1)
```

去除背景-形态学操作

膨胀

将结构元素的原点与二值图像中的1重叠，将二值图像中重叠部分不是1的值变为1，完成膨胀

A和B是两个集合，A被B膨胀定义为：

$$A \oplus B = \left\{ z \mid \left(\hat{B} \right)_z \cap A \neq \phi \right\}$$

腐蚀

将结构元素的原点覆盖在每一个二值图像的1上，只要二值图像上有0和结构元素的1重叠，那么与原点重叠的值为0

A和B是两个集合，A被B腐蚀定义为：

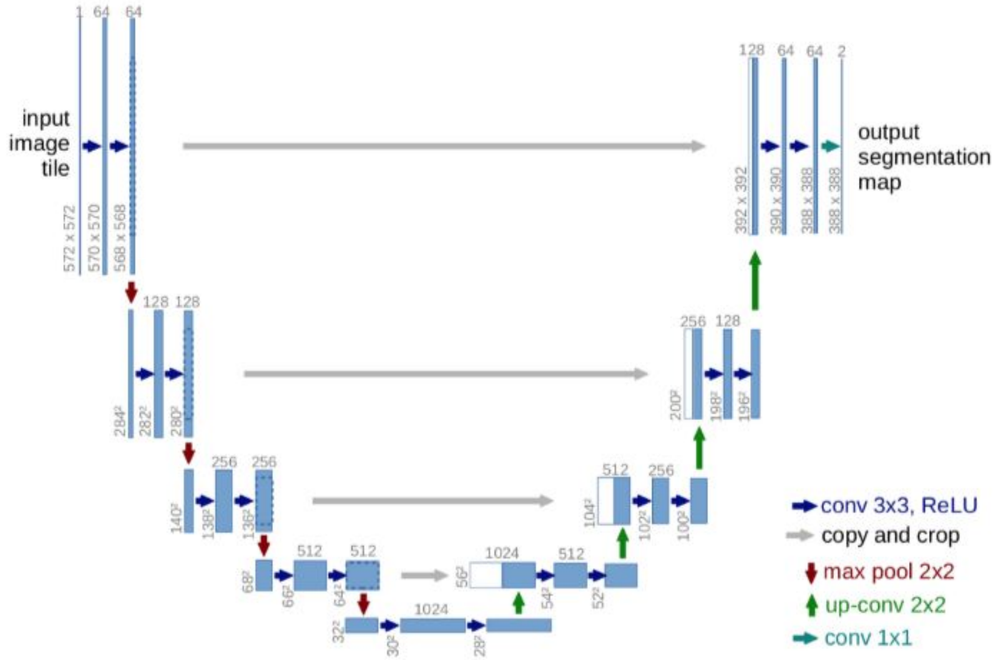
$$A \Theta B = \left\{ z \mid (B)_z \subseteq A \right\}$$

原始图像

腐蚀图像

膨胀图像

去除背景-Unet



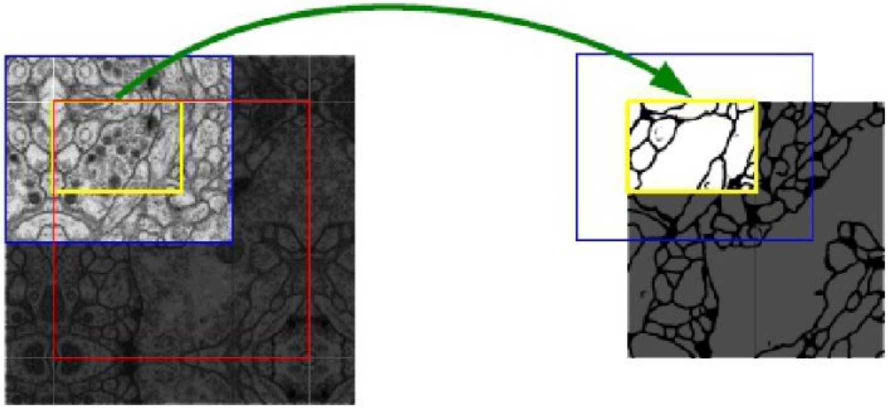
1.下采样。

2.提取特征。

3.上采样。

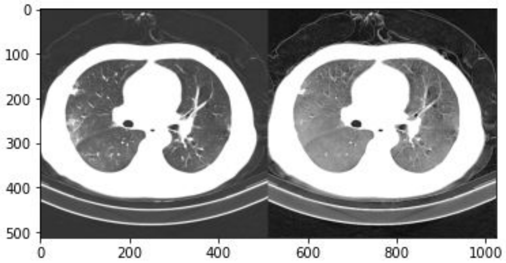
4.得到分割后的图像。

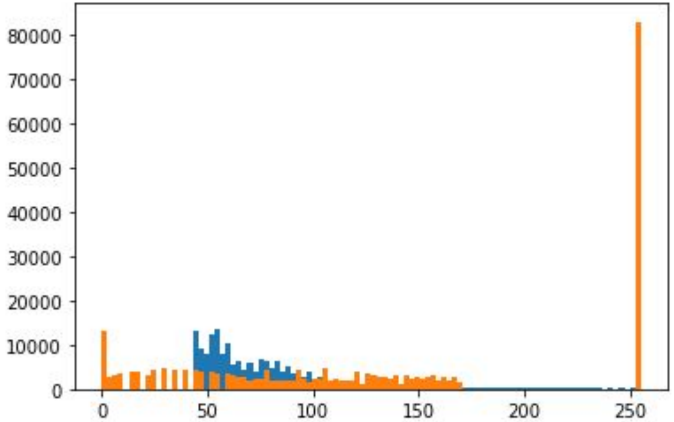
去除背景-Unet overlap-tile策略



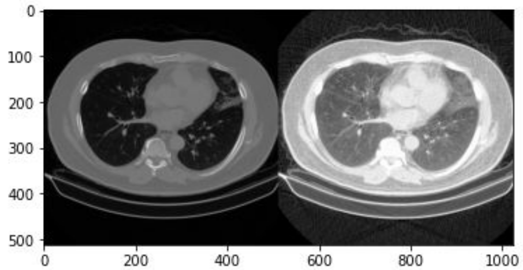
图像增强-HE

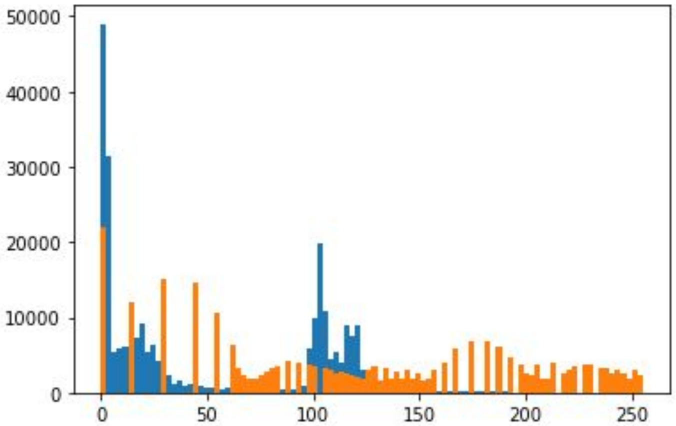
高对比度





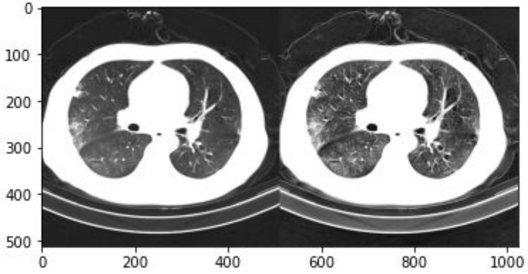
低对比度

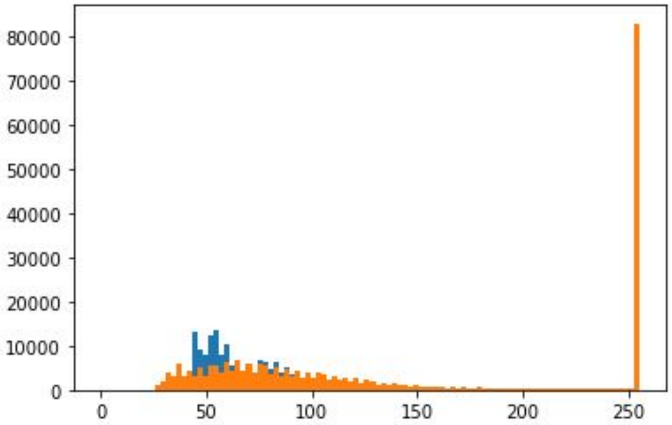




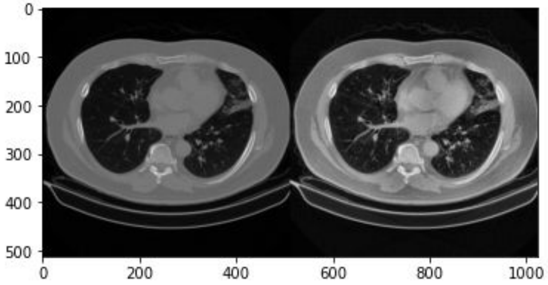
图像增强-CLAHE

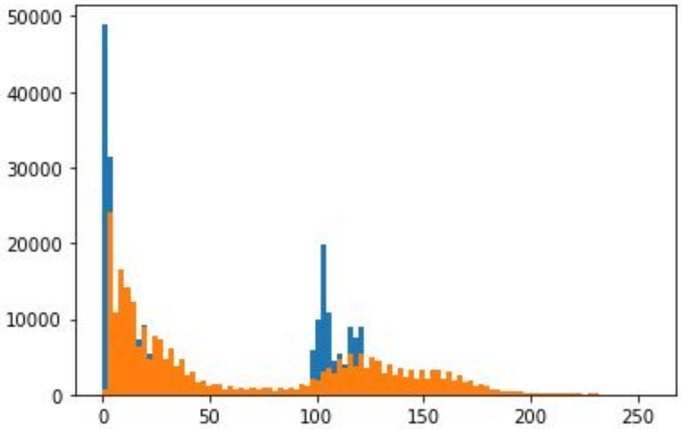
高对比度





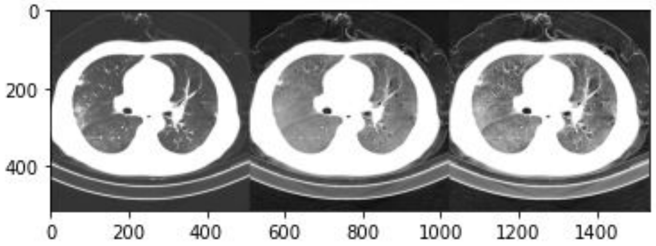
低对比度

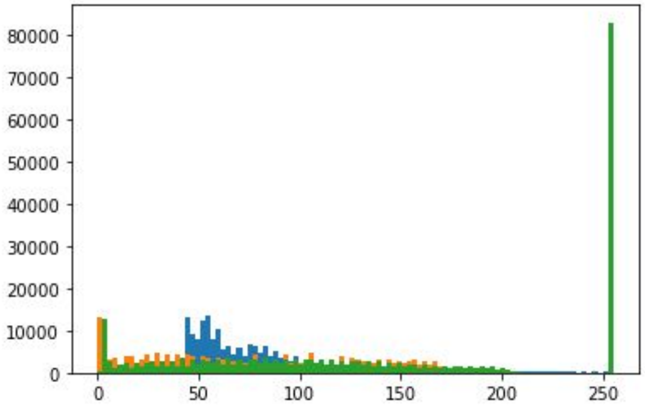




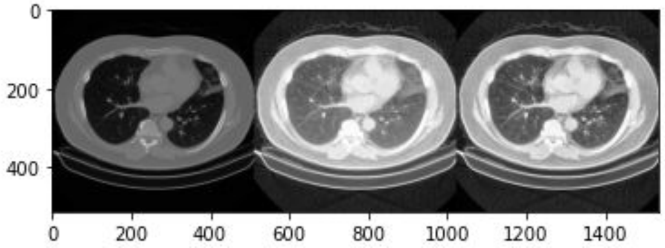
图像增强-HE+CLAHE

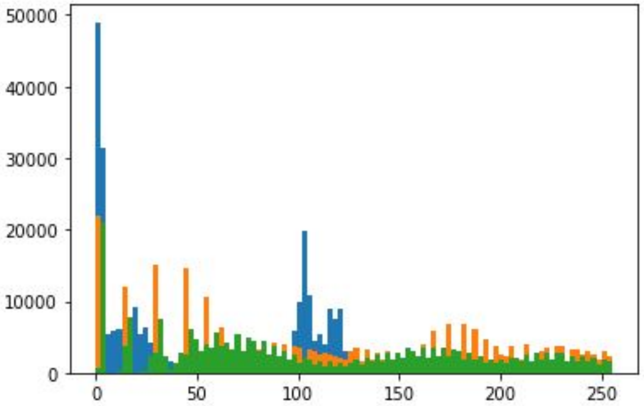
高对比度





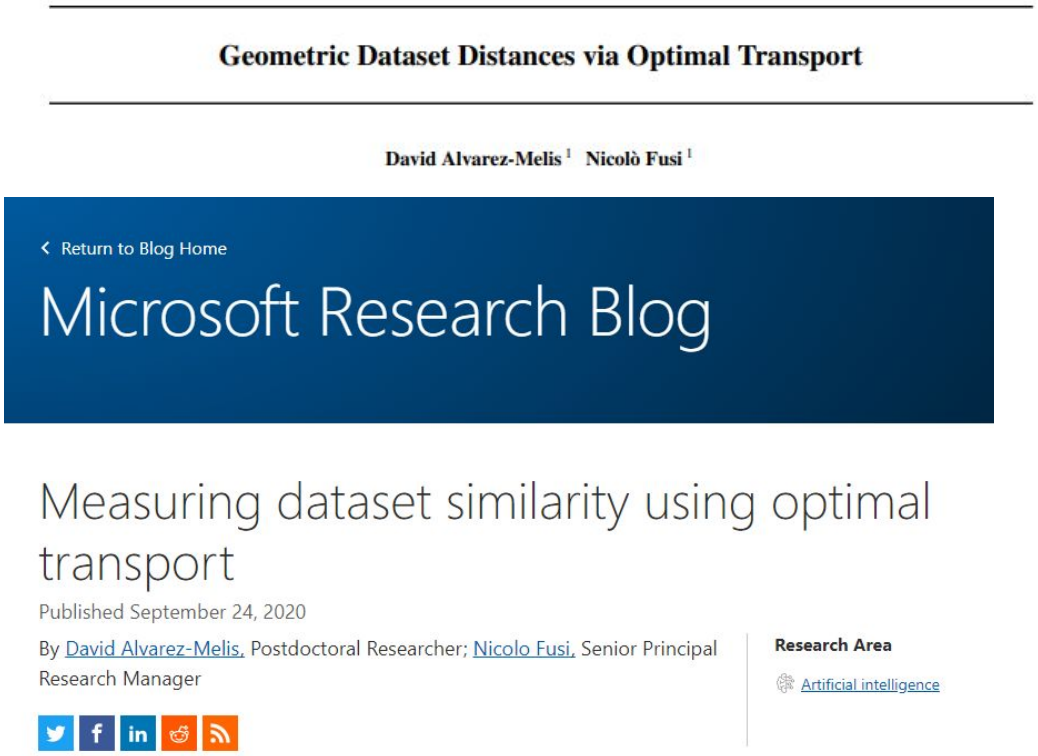
低对比度





OTDD

- 最佳传输的几何数据集距离 (OTDD)
  - 量化数据集之间的相似性
  - 以几何意义和原则性的方式比较特征标签对上的分布



Geometric Dataset Distances via Optimal Transport

David Alvarez-Melis [1] Nicolò Fusi [1]

‹ Return to Blog Home

Microsoft Research Blog

Measuring dataset similarity using optimal transport

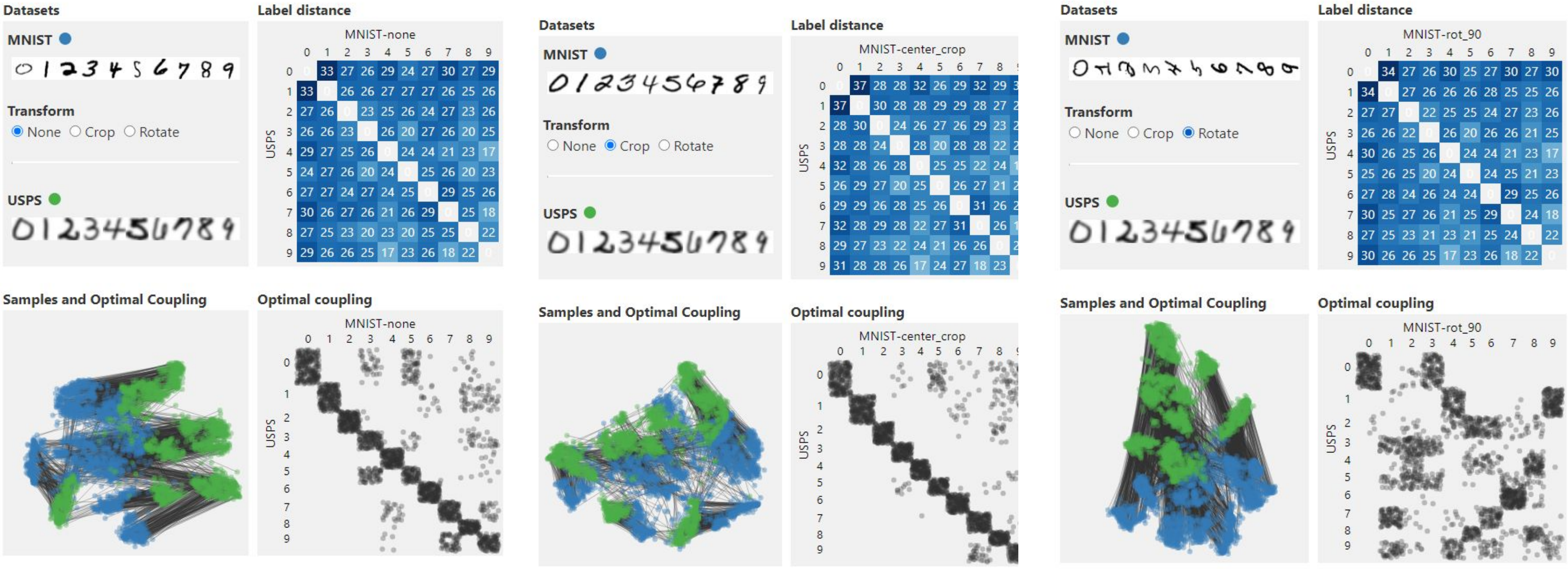
Published September 24, 2020

By David Alvarez-Melis, Postdoctoral Researcher; Nicolo Fusi, Senior Principal Research Manager

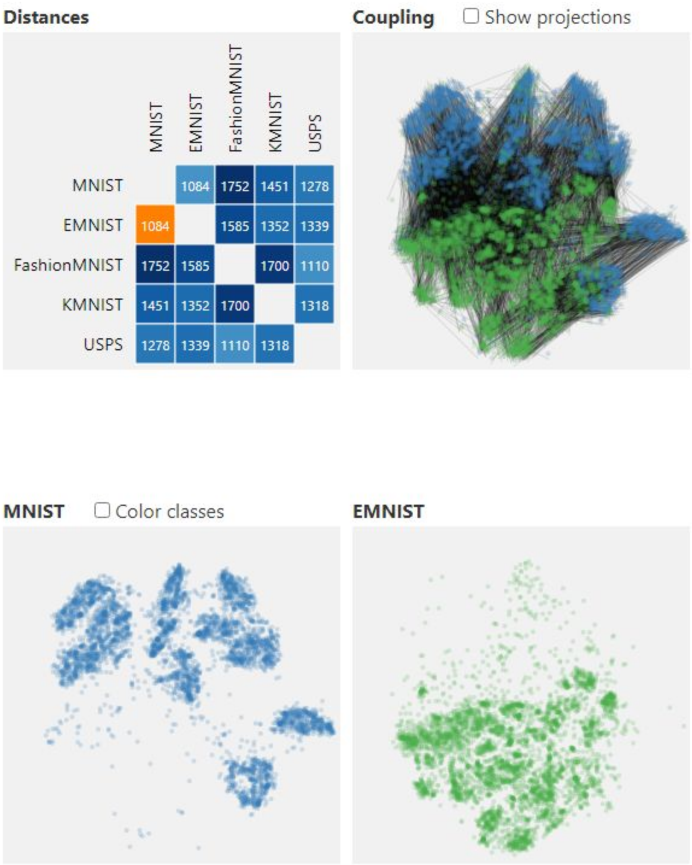
Research Area

Artificial intelligence

OTDD-指导数据增强

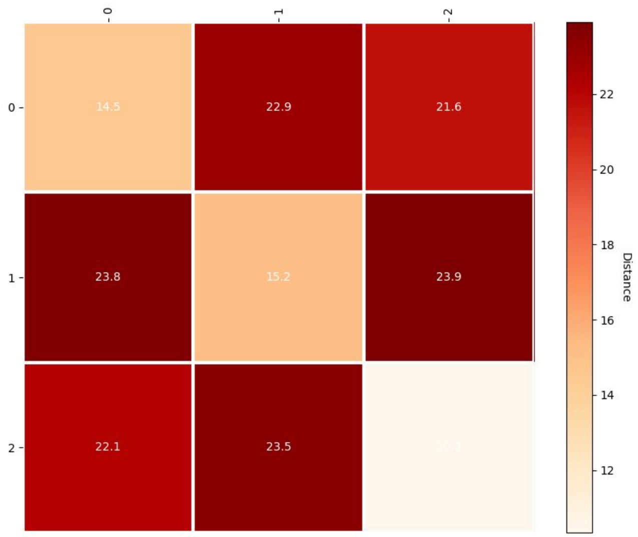

Datasets
MNIST
Transform
None
Crop
Rotate
USPS
Label distance
MNIST-none
USPS
Samples and Optimal Coupling
Optimal coupling
MNIST-none
Datasets
MNIST
Transform
None
Crop
Rotate
USPS
Label distance
MNIST-center_crop
Samples and Optimal Coupling
Optimal coupling
MNIST-center_crop
Datasets
MNIST
Transform
None
Crop
Rotate
USPS
Label distance
MNIST-rot_90
Samples and Optimal Coupling
Optimal coupling
MNIST-rot_90

OTDD-指导预训练


Distances
Coupling
Show projections
MNIST EMNIST FashionMNIST KMNIST USPS
MNIST 1084 1752 1451 1278
EMNIST 1084 1585 1352 1339
FashionMNIST 1752 1585 1700 1110
KMNIST 1451 1352 1700 1318
USPS 1278 1339 1110 1318
MNIST
Color classes
EMNIST

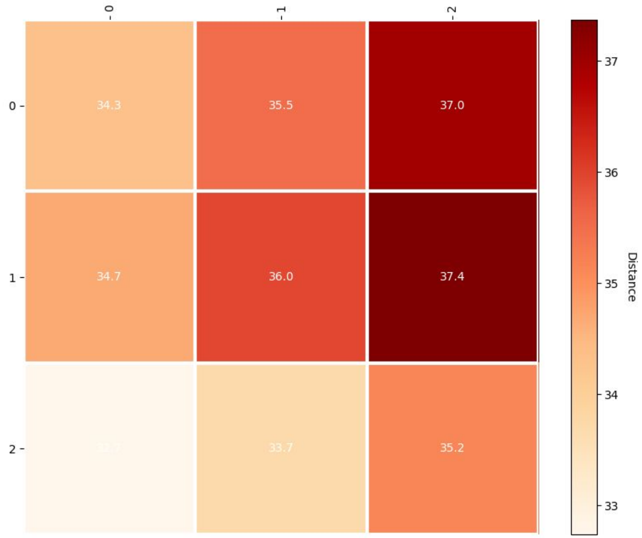
OTDD

- 直方图均衡结果与原始图像类间特征距离:



OTDD

- 高对比度与低对比度图像类间特征距离：



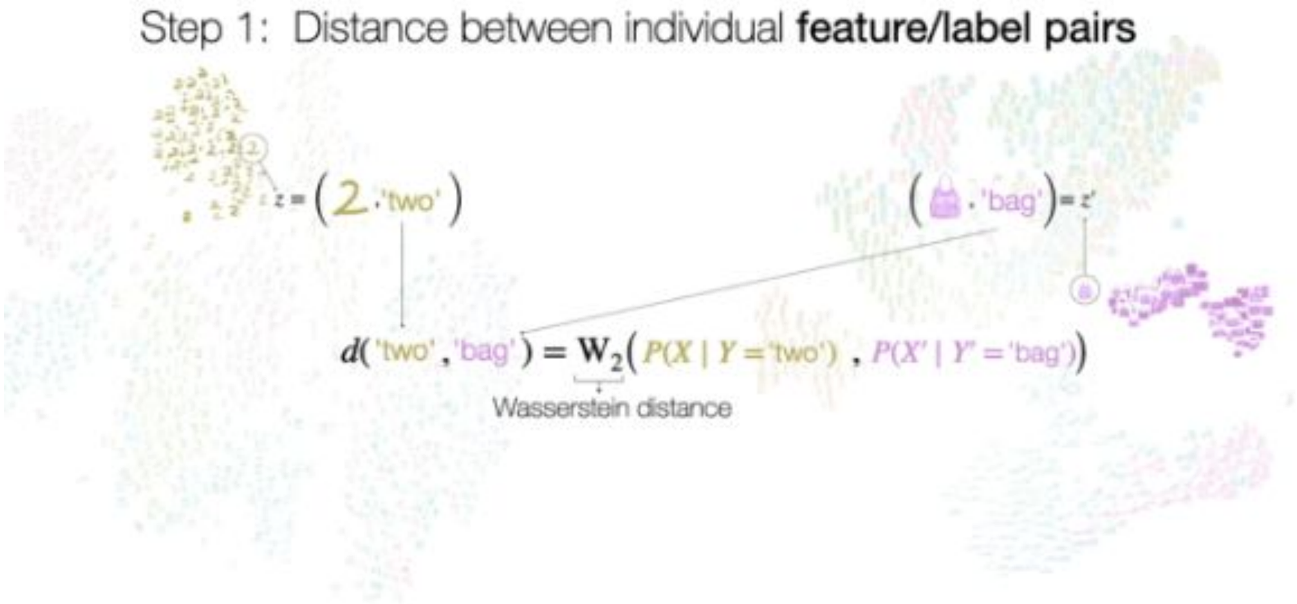
OTDD

Dataset A: Digits

How similar are they?

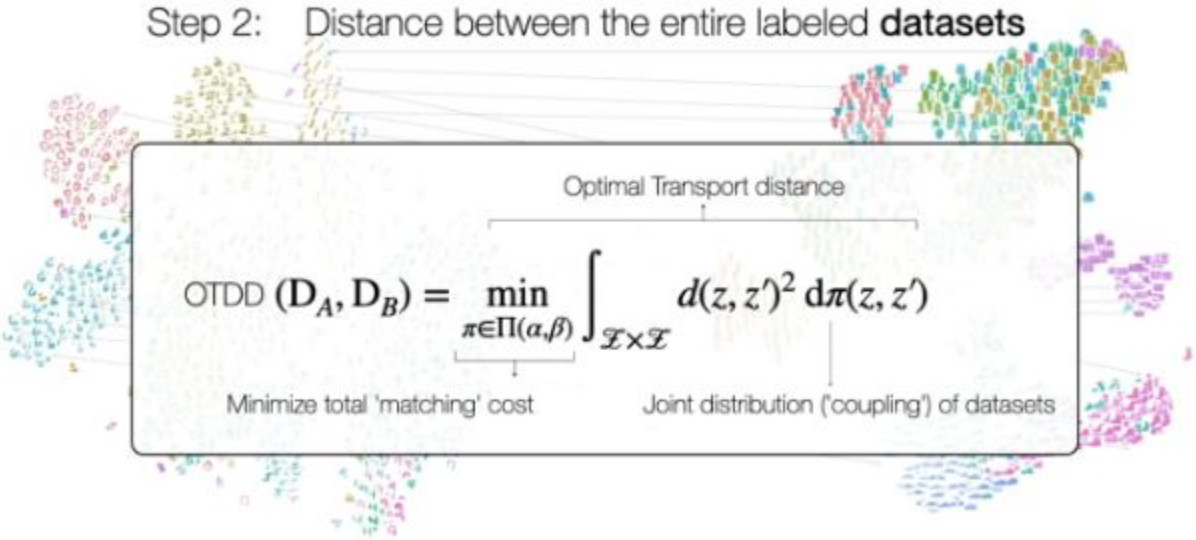
Dataset B: Clothes

OTDD



Step 1: Distance between individual **feature/label pairs**
z = (2, 'two')
(, 'bag') = z'
d('two', 'bag') = W2(P(X | Y = 'two') , P(X' | Y' = 'bag'))
Wasserstein distance

OTDD


Step 2: Distance between the entire labeled **datasets**
Optimal Transport distance
OTDD $(\mathrm{D}_A, \mathrm{D}_B) = \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{Z}\times\mathcal{Z}} d(z,z')^2 \, \mathrm{d}\pi(z,z')$
Minimize total 'matching' cost
Joint distribution ('coupling') of datasets

三、架构设计的变革

3.1 第一阶段的思路

Ensemble 大法好

决策融合是数据融合的一种形式，它将多个分类器的决策结合成一个共同决策。（软投票、硬投票）

Slice-Level -> 一级分类器+二级分类器

Subject-Level：端到端模型

Mangai U G, Samanta S, Das S, et al. A survey of decision fusion and feature fusion strategies for pattern classification[J]. IETE Technical review, 2010, 27(4): 293-307.

软投票 & 硬投票

软投票:以所有模型预测为某一类别的概率平均值作为标准投票。

硬投票:少数服从多数。

发现的问题

- 序列长度不同
    - 插值
        - 四种插值方式进行试验
        - 一二级分类器使用
    - 下采样
        - C3D（成本）
        - P3D
        - 各种端到端、视频理解
    - 长度不敏感结构
        - FCN
        - COVNet（maxpooling）
        - ConvLSTM

医学影像中的各向异性

一般来说，指的是CT、MR、PET这样的断层扫描数据中，x，y，z三个扫描方向的像素 间距（pixel spacing）不一致。为了便于后续分析，如卷 积等，一般在 预处理阶段通过图像插值把像素间距调整为各向同性（isotropic），也就是xyz三个方向的像素 间距一致。这里存在两种方式，上采 样或下采 样。

所谓上采样，就是把像素 间距统一向最小 间距（最高分辨率）靠 拢。比如把采 样后的CT像素间距设定为与xy方向一致。上采 样能够最大程度保留 图像信息，但生成的数据量大，对后续处理计算造成压力。下采 样则反之，就是把像素 间距统一向最大 间距（最低分辨率）靠 拢，比如，把采样后的CT像素间距设定为与z方向一致。下采 样会损失不少图像信息，但生成的数据量小.

Kurt B, Nabiyev V V, Turhan K. Medical images enhancement by using anisotropic filter and clahe[C]//2012 International Symposium on Innovations in Intelligent Systems and Applications. IEEE, 2012: 1-4.

医学影像中的各向异性

Preprocessing: CT images were extracted from Digital Imaging and Communications in Medicine (DICOM) files. We then followed the following steps to preprocess the CT images. First, the lung region is extracted as the region of interest (ROI) using a U-net (1) based segmentation method. Afterward, the extracted lung ROI was resampled to the same spacing (1 mm) in the z-direction. To reduce the dimensionality of the CT scan, we further down sampled the lung ROI by 5 times in the z-direction and scaled it to S × 224 × 224 pixels with S as the number of CT slices in the down sampled lung ROI. Then, the voxel intensity values were first clipped using a window/level of 1500HU/-500HU and were then normalized to the range of [0, 1]. To reduce the influence of vascular structure and boost the signals of the lesions, a maximum intensity projection (MIP) algorithm was also applied to each of the slices. Finally, the preprocessed image is then passed to the proposed COVID-19 detection neural network (COVNet) for the predictions.

Kurt B, Nabiyev V V, Turhan K. Medical images enhancement by using anisotropic filter and clahe[C]//2012 International Symposium on Innovations in Intelligent Systems and Applications. IEEE, 2012: 1-4.

3.2 第二阶段的思路

发现的问题

- 类别不均衡
    - Focal Loss
        - 效果不佳
    - Smote+ENN
        - 不适合图像
        - 适合图表
    - Slice-level subject-level 分布不均衡
        - 伪标签
    - 数据利用不充分
        - k-fold

半监督学习 & 伪标签技术

·半监督学习

·了解结合标记数据和未标记数据如何改变学习行为

·标记数据稀缺或昂贵时，使用现成的非标记数据来改进监督学习任务。

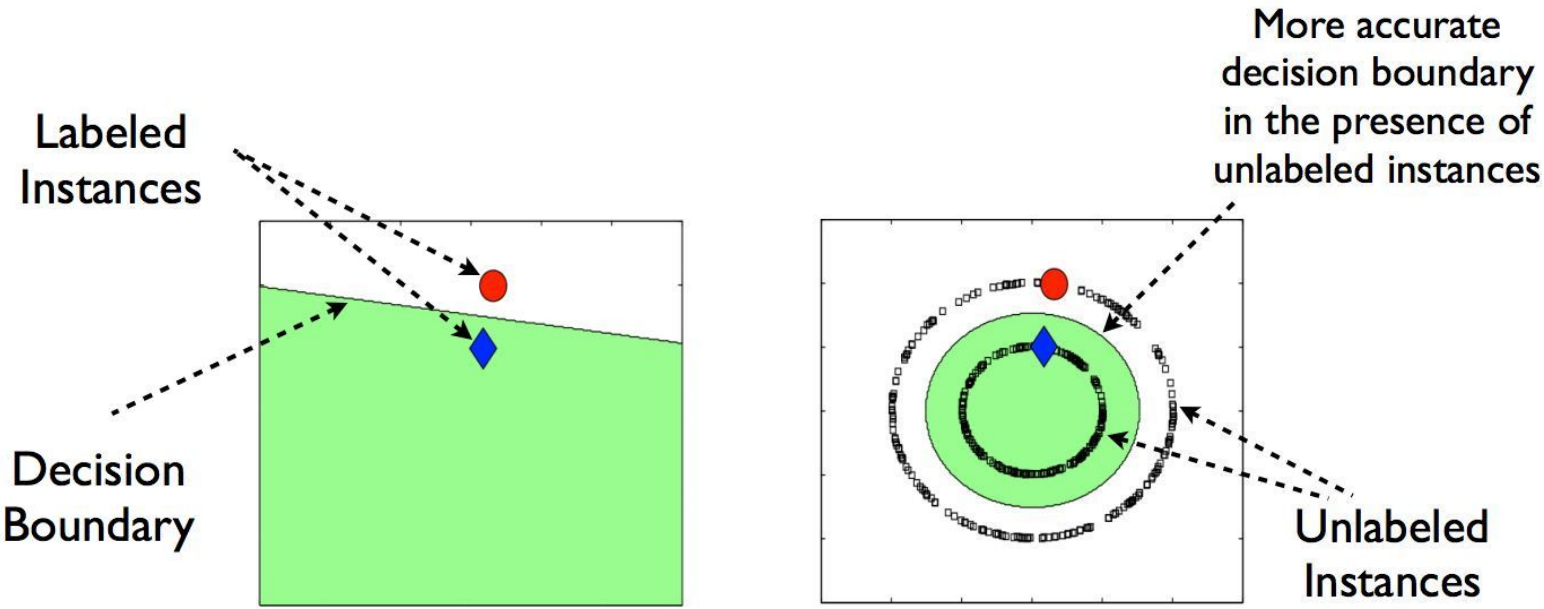
·伪标签技术

·利用在已标注数据所训练的模型在未标注的数据上进行预测，

Zhu X, Goldberg A B. Introduction to semi-supervised learning[J]. Synthesis lectures on artificial intelligence and machine learning, 2009, 3(1): 1-130.

Lee D H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks[C]//Workshop on challenges in representation learning, ICML. 2013, 3(2).

伪标签技术



采用无标签数据后，两个类别的决策边界变地更加精确了。

Focal Loss

- 给不同类别的样本加上权重。

$$\mathrm{CE}(p_t) = -\alpha_t \log(p_t).$$

Focal loss:

$$\mathrm{FL}(p_t) = -(1-p_t)^{\gamma} \log(p_t).$$

其中γ称为focusing parameter，γ≥0。

Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

注意力机制

神经注意力机制可以使得神经网络具备专注于其输入(或特征)子集的能力:选择特定的输入。

Zhao B, Wu X, Feng J, et al. Diversified visual attention networks for fine-grained object classification[J]. IEEE Transactions on Multimedia, 2017, 19(6): 1245-1256.

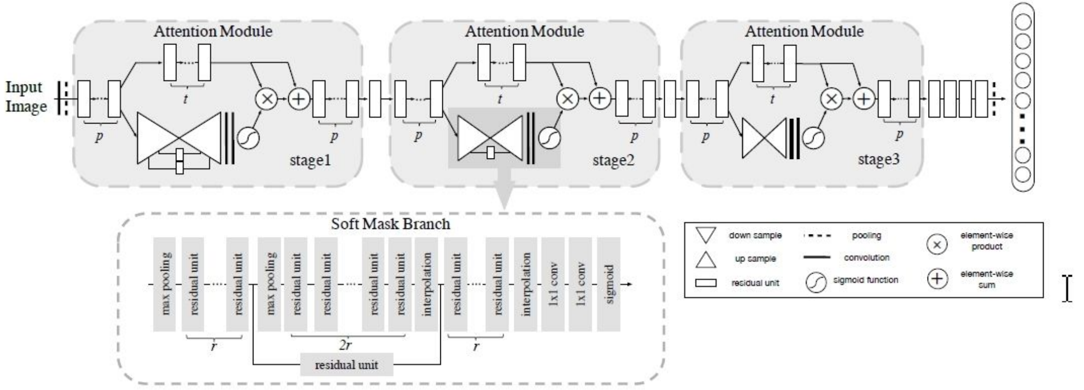
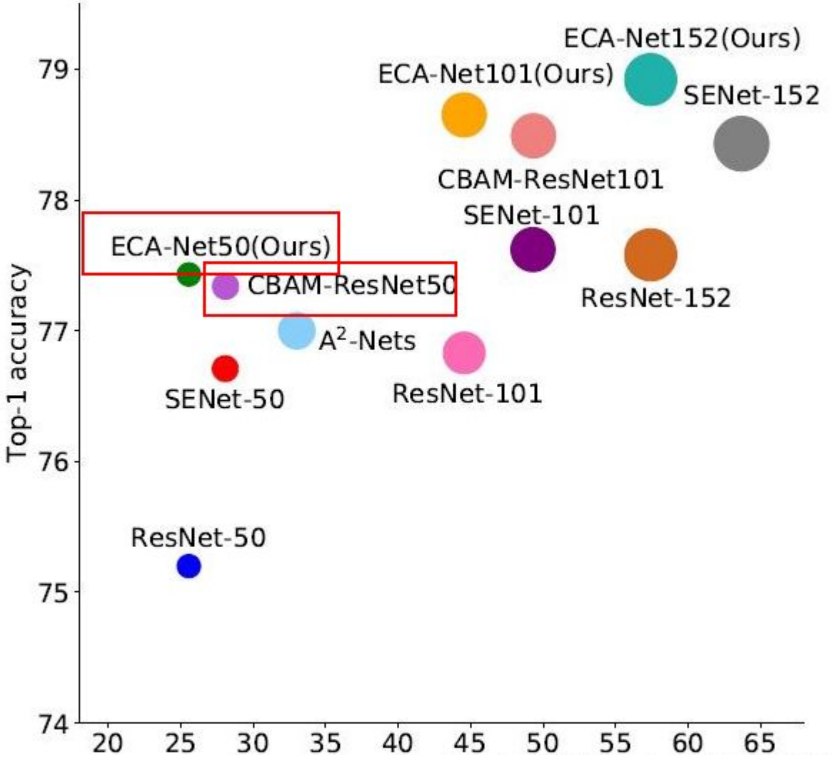


Figure 2: Example architecture of the proposed network for ImageNet. We use three hyper-parameters for the design of Attention Module: p, t and r. The hyper-parameter p denotes the number of pre-processing Residual Units before splitting into trunk branch and mask branch. t denotes the number of Residual Units in trunk branch. r denotes the number of Residual Units between adjacent pooling layer in the mask branch. In our experiments, we use the following hyper-parameters setting: $\{p = 1, t = 2, r = 1\}$. The number of channels in the soft mask Residual Unit and corresponding trunk branches is the same.

注意力机制



ECA-Net152(Ours)
ECA-Net101(Ours)
SENet-152
CBAM-ResNet101
SENet-101
ECA-Net50(Ours)
CBAM-ResNet50
ResNet-152
A2-Nets
SENet-50
ResNet-101
ResNet-50
Top-1 accuracy
79
78
77
76
75
74
20
25
30
35
40
45
50
55
60
65

3.3 第三阶段的思路

使用Gate

- 端到端中高低对比度效果差距极大
    - 低对比度效果极差

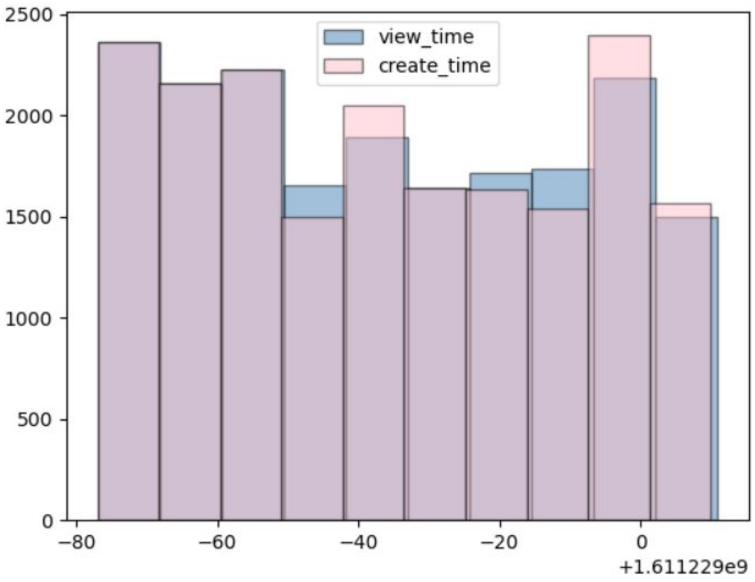
发现数据本身的思路

- 使用Gate:
  - 将数据按照高对比度和低对比度分成两类，分别输入到两个模型进行训练
  - 两个模型各自训练，预测时根据对比度分两个模型预测
-

Gate选择高低对比度

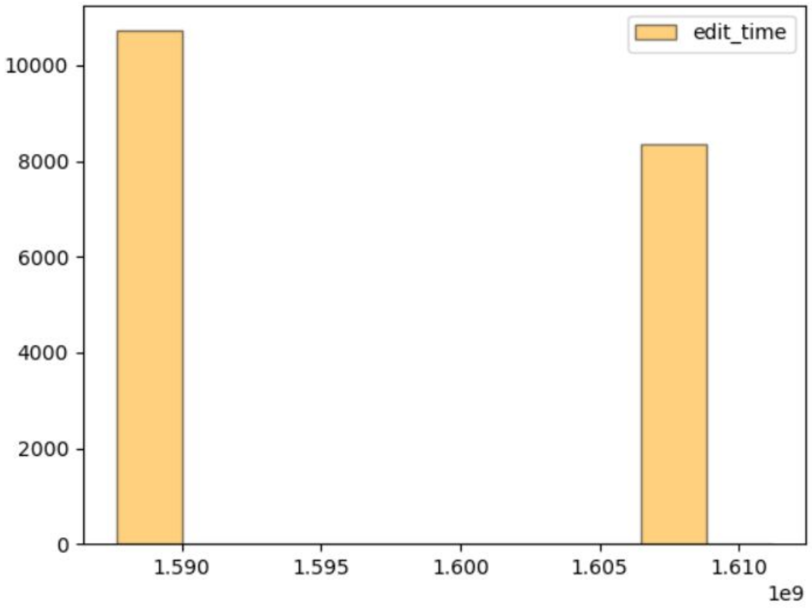
- 高对比度 端到端模型（COVNet+CSRNet作为backbone）
- 低对比度 一二级（一级（SelfTrans）+二级（SVM+Xgboost+软投票））

发现数据本身的分布

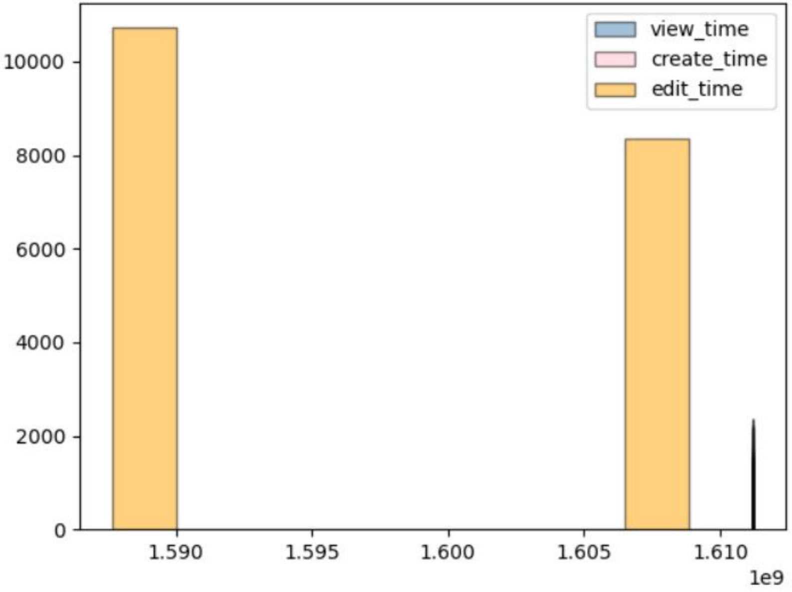


访问时间, 创建时间

发现数据本身的分布



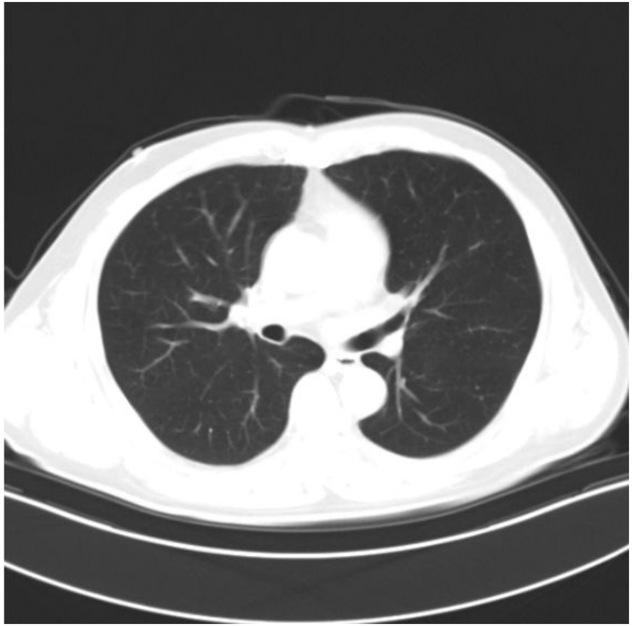
修改时间



访问时间 创建时间 编辑时间

发现数据本身的分布

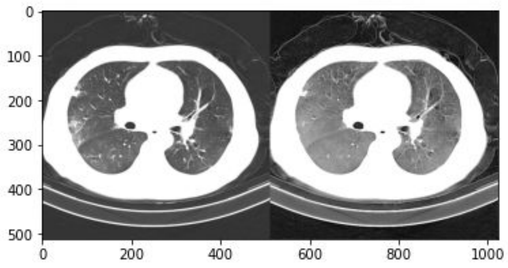
Low contrast

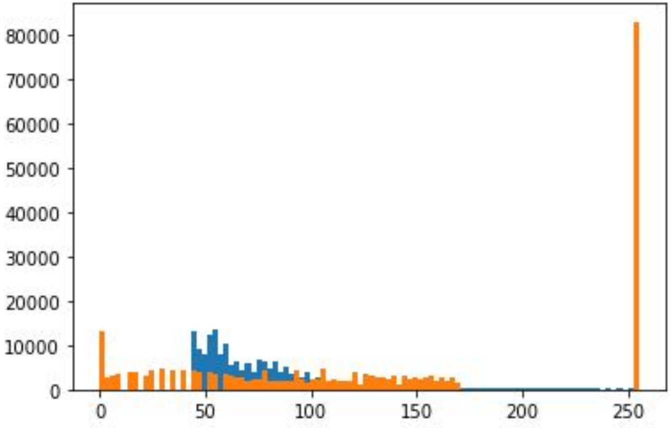


High contrast

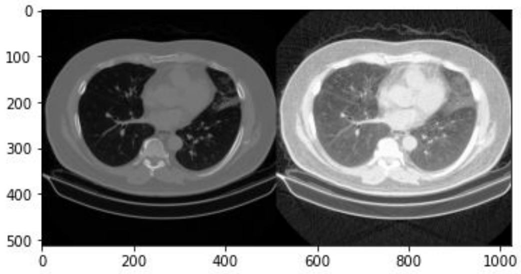
使用Gate

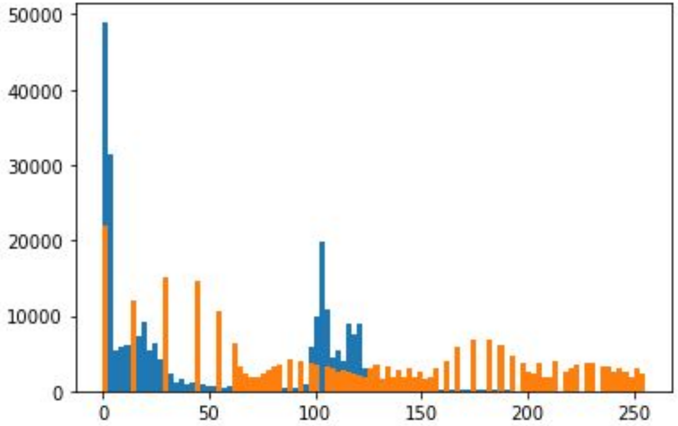
高对比度





低对比度





使用Gate-具体实现

```python
def graygate(I,threshold):
    I=np.array(I.convert('L'))
    I = I.reshape(-1)
    p_max=np.max(I)
    i=0
    for p in I:
        if p==p_max:
            i+=1
    return int(i>threshold)
```

网络结构

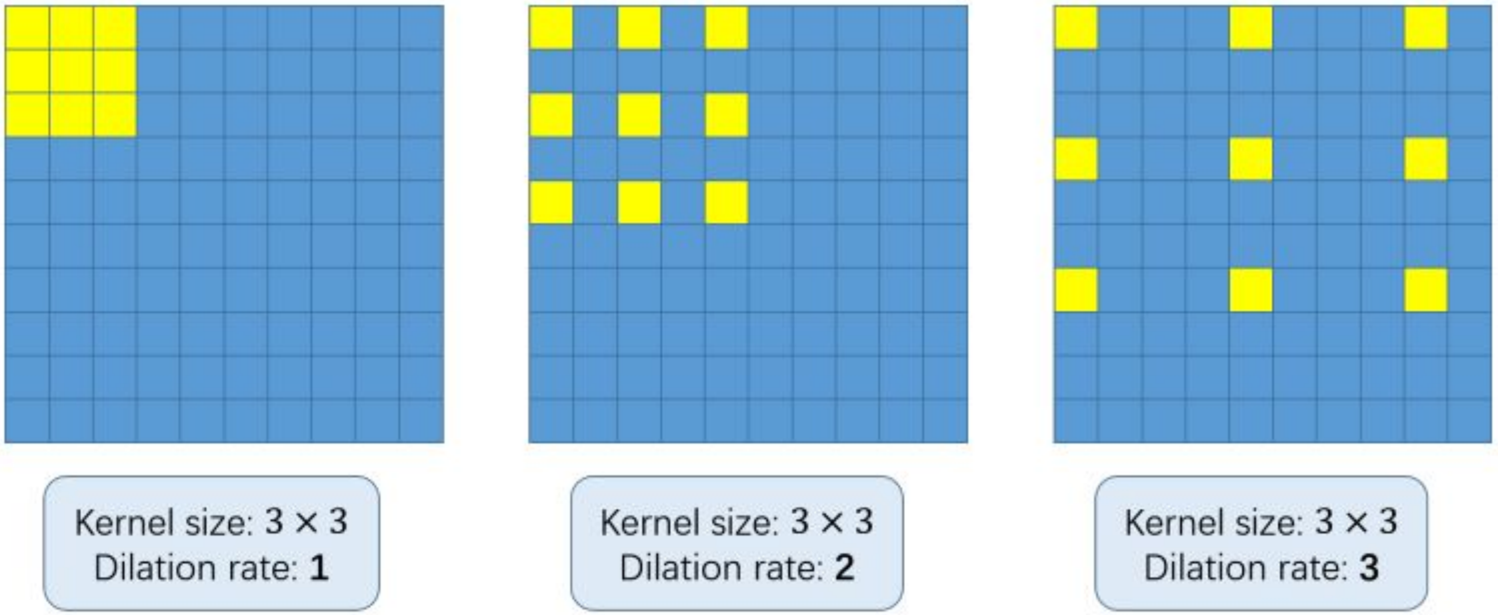
一级分类

- 01_pure_resnet
- 02_densenet_selftrans
- 03_COVNet
- 03_senet
- 04_resnet_plus

端到端

- 01_COVNet-50
- 02_COVNet-18
- 03_COVNet-34
- 04_COVNet-50-dropout-only
- 05_COVNet-FCN
- 06_COVNet-CSR
- 06_COVNet-CSR_inf
- 07_COVNet-50-CBAM
- 08_COVNet-50-ECA
- 09_COVNet-CSR-Gate
- 10_COVNet-CSR-Focal
- 11_COVNet-CSR-Focal-CBAM

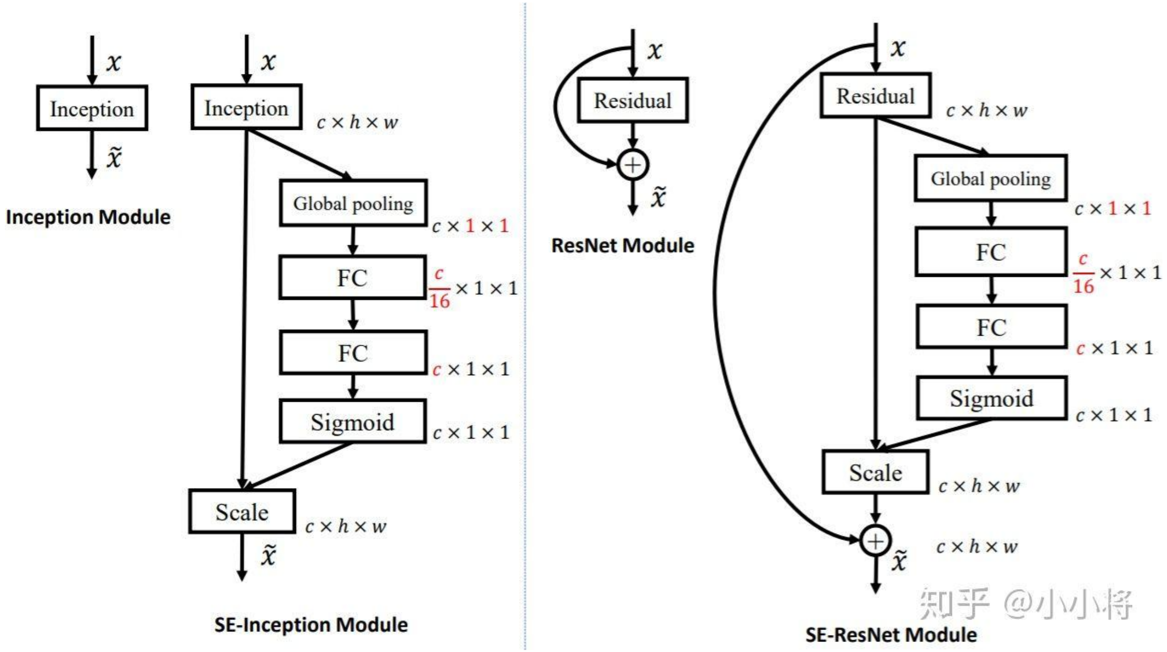
CSRnet


Kernel size: 3 × 3
Dilation rate: **1**
Kernel size: 3 × 3
Dilation rate: **2**
Kernel size: 3 × 3
Dilation rate: **3**

CSRnet

Configurations of CSRNet			
A	B	C	D
input(unfixed-resolution color image)			
front-end (fine-tuned from VGG-16)			
conv3-64-1 conv3-64-1			
max-pooling			
conv3-128-1 conv3-128-1			
max-pooling			
conv3-256-1 conv3-256-1 conv3-256-1			
max-pooling			
conv3-512-1 conv3-512-1 conv3-512-1			
back-end (four different configurations)			
conv3-512-1 conv3-512-1 conv3-512-1 conv3-256-1 conv3-128-1 conv3-64-1	conv3-512-2 conv3-512-2 conv3-512-2 conv3-256-2 conv3-128-2 conv3-64-2	conv3-512-2 conv3-512-2 conv3-512-2 conv3-256-4 conv3-128-4 conv3-64-4	conv3-512-4 conv3-512-4 conv3-512-4 conv3-256-4 conv3-128-4 conv3-64-4
conv1-1-1			

Senet:Resnet加入注意力机制



Densenet

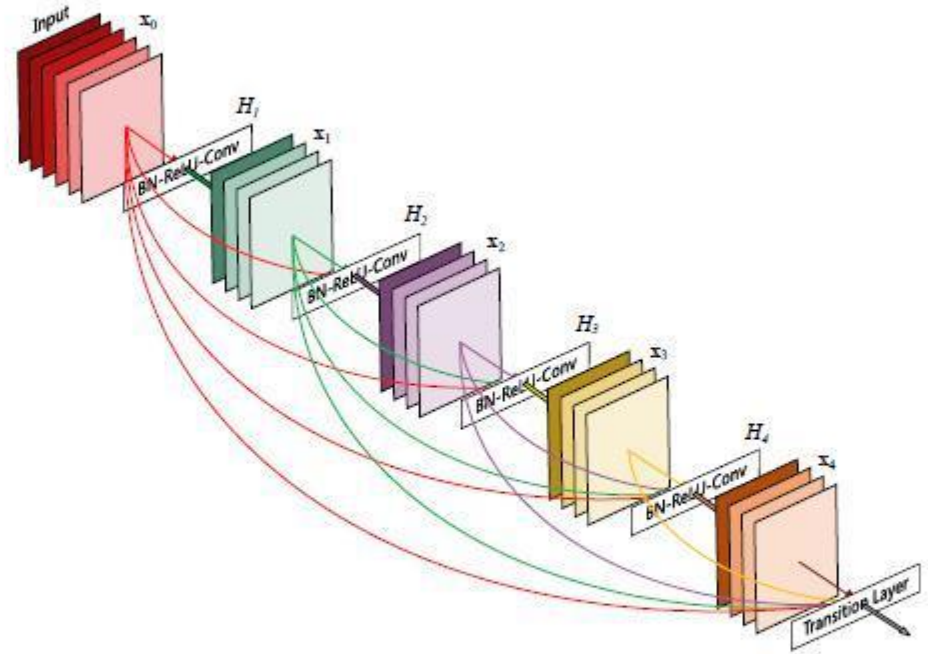


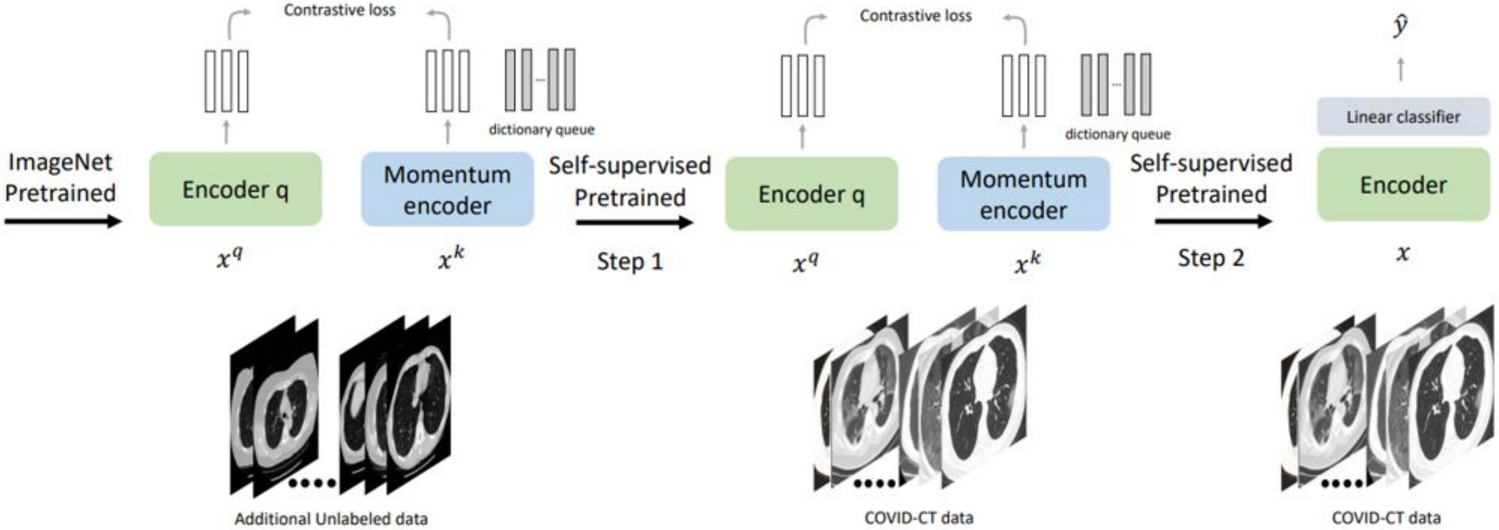
Figure 1. A 5-layer dense block with a growth rate of $k = 4$. Each layer takes all preceding feature-maps as input.

Self-Trans(对比自监督学习)

- **存在问题:**
  - 数据集规模小
  - 存在过拟合风险

- **原理:**
  - 预训练:利用数据丰富的源任务来帮助学习数据不足的目标任务
  - 自监督学习:减少模型对源图像及其类标签的偏差，提高对目标数据的泛化效果

Self-Trans(对比自监督学习)



Self-Trans(对比自监督学习)

- **步骤:**
  - 在ImageNet数据集上进行预训练
  - 在肺结节恶性度(LNM)数据集上进行自监督学习
  - 在目标数据集上进行自监督学习
  - 微调网络，通过目标数据集进行有监督学习
- **自监督学习:**
  - 目的:使用不带标签的数据，学习输入数据的有意义的表示
  - 辅助任务**:**判断通过随机数据增强生成的两幅图像是否是同一幅原始图像的增强
- 网络:DenseNet-169

二级分类器

- SVM+XGBoost
- Grid Search

	C = 0.001	C = 0.01	...	C = 10
gamma=0.001	SVC(C=0.001, gamma=0.001)	SVC(C=0.01, gamma=0.001)	...	SVC(C=10, gamma=0.001)
gamma=0.01	SVC(C=0.001, gamma=0.01)	SVC(C=0.01, gamma=0.01)	...	SVC(C=10, gamma=0.01)
...	...	...	...	...
gamma=100	SVC(C=0.001, gamma=100)	SVC(C=0.01, gamma=100)	...	SVC(C=10, gamma=100)

四、神经网络可解释性

中间层特征值

CSR_GATE_HIGHmodel_best

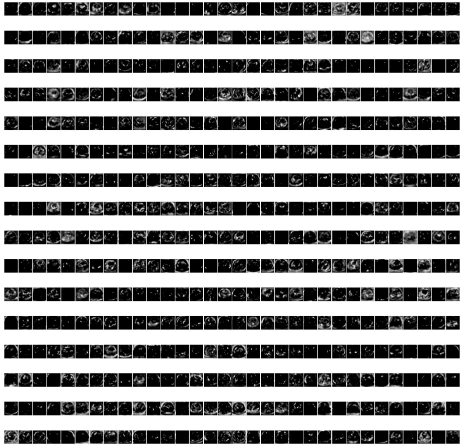
CSR_HE_Intepmodel_best

CSR_HIGH_Focalmodel_best

中间层特征值

CSR_GATE_HIGHmodel_best

CSR_HE_Intepmodel_best

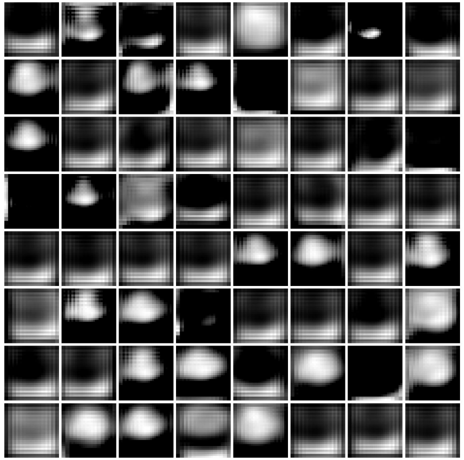


CSR_HIGH_Focalmodel_best

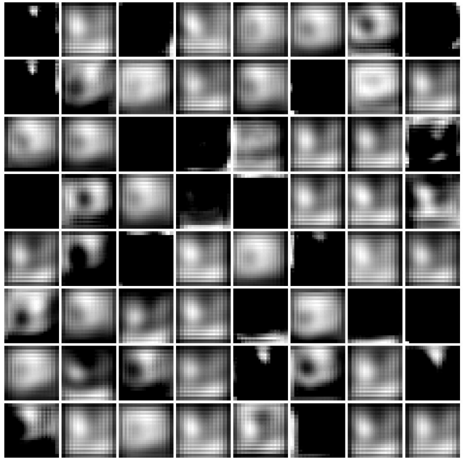
中间层特征值

CSR_GATE_HIGHmodel_best

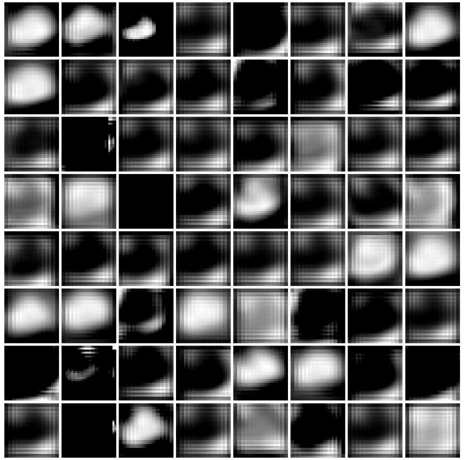
中间层特征值



CSR_GATE_HIGHmodel_best

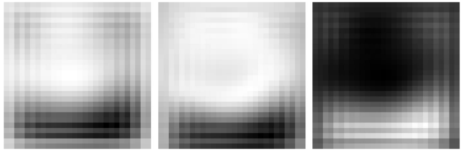


CSR_HE_Intepmodel_best

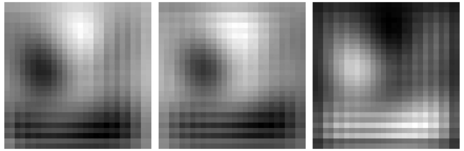


CSR_HIGH_Focalmodel_best

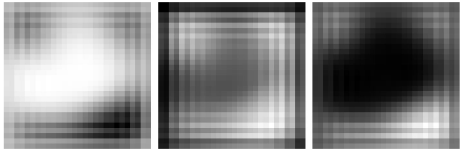
中间层特征值



CSR_GATE_HIGHmodel_best



CSR_HE_Intepmodel_best



CSR_HIGH_Focalmodel_best

训练集上结果

test done

	precision	recall	f1-score	support
Normal	0.92	1.00	0.96	12
CAP	1.00	0.92	0.96	13
COVID-19	1.00	1.00	1.00	20
accuracy			0.98	45
macro avg	0.97	0.97	0.97	45
weighted avg	0.98	0.98	0.98	45

Low Contrast

Test Report

	precision	recall	f1-score	support
0	0.94	1.00	0.97	16
1	0.94	0.94	0.94	16
2	0.90	0.82	0.86	11
accuracy			0.93	43
macro avg	0.93	0.92	0.92	43
weighted avg	0.93	0.93	0.93	43

High Contrast

工具链

Pytorch Sklearn

数据挖掘与分析的简单而有效的工具。大家都懂orz

WandB

wandb是Weights & Biases的缩写，能够帮助跟踪机器学习项目

4项核心功能：

- 看板：跟踪训练过程，给出可视化结果
- 报告：保存和共享训练过程中一些细节、有价值的信息
- 调优：使用超参数调优来优化你训练的模型
- 工具：数据集和模型版本化

感谢聆听！

祝大家身体健康,万事如意!

寒假快乐!