Evaluating Grad-CAM heatmaps of X-ray images

Vaggelis Lamprou

Presentation

Contents

1. Data

- COVID-19 Radiograhy Database (link)
- Classes: Covid, Lung Opacity, Normal, Viral Pneumonia
- Down-sampling: 1345 images per class
- Training set (80% per class): 1076 * 4 = 4304 images
  Validation set (10% per class): 134 * 4 = 536 images
  Test set (10% per class): 135 * 4 = 540 images

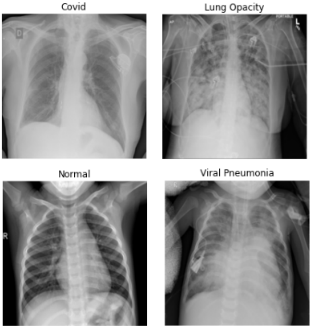


Figure: X-ray examples

2. Classifier Models

- Transfer Learning approach

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 vgg16 (Functional)          (None, 9, 9, 512)         14714688

 max_pooling2d (MaxPooling2D  (None, 2, 2, 512)        0
 )

 flatten (Flatten)           (None, 2048)              0

 dense (Dense)               (None, 64)                131136

 dense_1 (Dense)             (None, 4)                 260

=================================================================
Total params: 14,846,084
Trainable params: 131,396
Non-trainable params: 14,714,688
_________________________________________________________________
```

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 densenet201 (Functional)    (None, 9, 9, 1920)        18321984

 max_pooling2d_1 (MaxPooling  (None, 2, 2, 1920)       0
 2D)

 flatten_1 (Flatten)         (None, 7680)              0

 dense_2 (Dense)             (None, 64)                491584

 dense_3 (Dense)             (None, 4)                 260

=================================================================
Total params: 18,813,828
Trainable params: 491,844
Non-trainable params: 18,321,984
_________________________________________________________________
```

- Compiled with Categorical Crossentropy loss and Adam optimizer

3. The Grad-CAM algorithm

- Post-hoc
- Visual explanations
- Local explanations
- Denote by $A_1, A_2, .., A_K$ the activation maps of the last convolutional layer of the CNN-based model.
  Class dependent heatmaps are constructed as follows:

$$\underbrace{L^c_{Grad-CAM} = ReLU(\sum_k \underbrace{\left[\frac{1}{Z}\sum_i\sum_j \overbrace{\frac{\partial y^c}{\partial A^k_{ij}}}^{\text{step 1}}\right]}_{\text{step 2: } a^k_c} A^k)}_{\text{step 3}}$$

https://arxiv.org/pdf/1610.02391.pdf

4. Area Over Perturbation Curve

- How do we choose among two heatmaps ?
- Use MoRF technique to evaluate heatmaps.
  Looks at the heatmaps as a decreasing sequence of importance regions $\{r_1, r_2, .., r_L\}$. For image x, compute images sequence by:

$$x_{MoRF}^{(0)} = x$$
$$x_{MoRF}^{(k)} = g(x_{MoRF}^{(k-1)}, r_k), \quad k = 1, 2, .., L$$

  and consider the MoRF perturbations graph as defined by the points

$$\{(k, f(x_{MoRF}^{(k)})), \quad k = 0, 1, .., L\}.$$

- Denote by $< \cdot >$ the average over the entire test set

$$AOPC = \frac{1}{L+1} < \underbrace{\sum_{k=1}^{L} [f(x_{MoRF}^{(0)}) - f(x_{MoRF}^{(k)})]}_{\text{per image controls area over MoRF curve}} >$$

https://arxiv.org/pdf/1509.06321.pdf

5. Results: Train & Test

- The models were trained for 20 epochs and monitored with Model Checkpoint callback
- After training, get model instance that maximizes the diseases validation recall scores.
- Eventually, kept VGG16 based model of epoch 16 and DenseNet201 based model of epoch 15.

	precision	recall	f1-score	support
Covid	0.88	0.84	0.86	135
Lung Opacity	0.82	0.82	0.82	135
Normal	0.79	0.82	0.80	135
Viral Pneum	0.96	0.96	0.96	135
accuracy	-	-	0.86	540
macro avg	0.86	0.86	0.86	540
weighted avg	0.86	0.86	0.86	540

Figure 4: VGG16 based model - Classification report

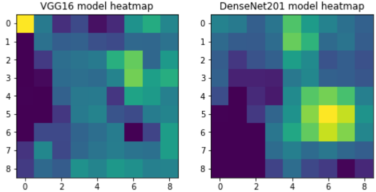
	precision	recall	f1-score	support
Covid	0.91	0.87	0.89	135
Lung Opacity	0.79	0.81	0.80	135
Normal	0.83	0.86	0.85	135
Viral Pneum	0.99	0.98	0.98	135
accuracy	-	-	0.88	540
macro avg	0.88	0.88	0.88	540
weighted avg	0.88	0.88	0.88	540

Figure 5: DenseNet201 based model - Classification report

- Note DenseNet201 has better recall and f1 scores! What about AOPC score ?

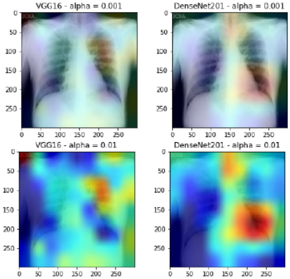
5. Results: Grad-CAM Heatmaps Visualization

- Consider Covid test image correctly classified by both models
- The heatmaps of the top predicted class are:



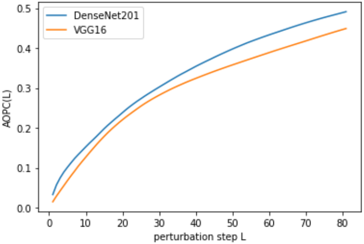
Note that they are divided in 81 regions (approx (34,34) pixel blocks)

- Super-imposed images on different α levels:



5. Results: AOPC graphs

- Per classifier compute AOPC values as function of perturbation step L, considering only correctly classified test images.
- Perturbations via random normal noise of mean 0 and standard deviation 0.1 (function g).



- Scores summary:

Base model	Test Recall	Test F1	AOPC
VGG16	86%	86%	0.45
DenseNet201	88%	88%	0.49

Thank You !