Vol.2019-CVIM-218 No.16
2019/9/5

Black Average Drop: 医用画像に対する可視化選択指標

下村 真生[†1,a)]　中村 和幸[†1,†2,b)]

概要：医療画像診断支援に最も適した可視化を選ぶ指標について議論する．深層学習を用いた医用画像分類において，判断根拠の可視化は医師の診断を支援する観点から重要である．可視化手法は複数あるが，データセットに適した手法選択に関する議論は不十分である．ピクセルラベルなしに測る既存指標 Average Drop は一般画像に対して有効だが，画像の不透明度を変更したもので測る等、医療画像には不適切であった。本研究では等積のマスクをした画像を用いて測る Black Average Drop を提案する。既存手法と異なり，CAM や Grad-CAM に比べ小病変を見逃す危険の少ない Grad-CAM++が医用画像に適するという正しい判断を得ることが提案手法を用いた評価によりできた．

1. はじめに

医療画像撮影技術の進展に伴い，胸部レントゲン写真の枚数は増加傾向にある．一方で，それを読影する放射線医の人数は不足しており，1 枚にかけられる時間が少なかったり，非専門医が診断したりしている [1]．その結果，病変の見逃しによる重大な医療事故が発生し，ニュースにも度々取り上げられてきた [2]．このような状況を変えるため，人工知能を用いた自動診断技術の開発がここ数年盛んである．いくつかの研究では，既に特定のデータセットに対する診断精度が専門医を超えたと報告しており，注目を集めている [3]．しかし，日本では厚生労働省より「人工知能（AI）を用いた診断・治療支援を行うプログラムを利用して診療を行う場合についても，診断，治療等を行う主体は医師であり，医師はその最終的な判断の責任を負う」との発表が昨年公開された [4]．臨床医が本当に求めるツールとしての自動診断技術とは，完璧な精度で診断するだけでなく，診療の中で使うに匹敵するだけのエビデンスを持ったツールである．そこで，医師がどこを見たら正しく診断できるかを可視化するという研究も始まっている．

Deep Learning を用いた画像分類において，その判断根拠を可視化する技術は guided backpropagation[5] や SmoothGrad[6] など複数存在する．しかし，それらの多くはノイズが多く，臨床医が見て解釈をするには少々難しい．したがって，比較的わかりやすい CAM[7] や Grad-CAM[8] などが X 線画像に対し頻繁に使用されている．近頃，Grad-CAM を改良した Grad-CAM++[9] が登場するなど，可視化手法は増加傾向にあり，今度は適用したいデータセットに対する手法選択が必要となってきている．Semantic Segmentation の分野では，ピクセル単位のラベリングがされたデータに対しては pixel-accuracy や mean Intersection over Union (mIoU) などの評価指標が存在する．しかし，医療画像などの分野ではピクセル単位のラベリングをすることは非常に困難で，そのようなデータを大量に含むデータセットは入手が難しい．したがって，ピクセル単位のラベルがない場合でも使用できる評価指標が必要であるが，その点に関する議論は不十分である．Chattopadhyay ら [9] が提案した Average Drop (AD) は，入力画像とそれを加工した画像をそれぞれ Convolutional Neural Network (CNN) に通したときのスコアの差を元に評価を行う．その加工画像は，CAM によって描いたヒートマップのもつピクセル単位の重要度を画像のアルファ値に反映させて作成する．この指標は ImageNet に含まれるような一般的な画像に対しては有効であるが，X 線画像には不適切である．放射線医は X 線写真を見るとき，正常部より透過性が亢進または低下している箇所を見つけることによって診断を行う．その重要な情報を，その加工画像は損なっているのである．

本研究報告では，Deep Learning の二値分類問題を用いた肺炎の診断支援に焦点を当てる．まず，CAMs[*1]を用い

[†1] 現在，明治大学
Presently with Meiji University
[†2] 現在，国立研究開発法人科学技術振興機構，さきがけ
Presently with JST, PRESTO
[a)] dason.data@gmail.com
[b)] knaka@meiji.ac.jp

[*1] 本研究では，Class Activation Mapping (CAM)，Grad-CAM，Grad-CAM++をまとめて CAMs (CAM series) と呼ぶ．

た疾病部位の可視化を行い，Grad-CAM++が医療画像には最適であることの説明を行う．その後，可視化に対する新しい評価手法である Black Average Drop の提案を行う．これは医療画像へ適用時に生じる AD の欠点を解決するために，アルファ値を変化させないようにするなどの改良を行っている．その結果，Grad-CAM++が胸部レントゲン写真に対して最善であるという正しい評価をすることが，この指標ではできた．

2. 先行研究

2.1 Class Activation Mapping

Class Activation Mapping（CAM）[7] の利用には，CNN に画像を入力し，最後の Convolution 層の出力として得られる特徴量マップと，それに対して Global Average Pooling を施し，その k 番目のノード (特徴量マップの k 番目のチャンネルに対応) とクラス c の出力への重みを使用する．

まず，A^k_{xy} は特徴量マップの k 番目のチャンネル の位置 (x, y) の活性を表すとする．さらに，Global Average Pooling[10] を施した結果として F^k は次の計算される．

$$F^k = \frac{1}{z}\sum_x\sum_y A^k_{xy} \tag{1}$$

ここで，z は特徴量マップの面積 $(x \times y)$ である．この後，最後の分類層への入力 y^c は次の計算によって得られる．

$$y^c = \sum_k w^c_k \frac{1}{z}\sum_x\sum_y A^k_{xy}\left(\sum_k w^c_k F^k\right) \tag{2}$$

ここで w^c_k は特徴量マップの各チャンネルのクラス c に対する重要度を表す．この式は次のように変形できる．

$$y^c = \sum_x\sum_y\left(\frac{1}{z}\sum_k w^c_k A^k_{xy}\right) \tag{3}$$

さらに，L^c_{xy} を入力画像の位置 (x, y) のクラス c への活性を表すとすると，式 (3) は $L_{xy} = \frac{1}{z}\sum_k w^c_k A^k_{xy}$ を用いて

$$y^c = \sum_x\sum_y L^c_{xy} \tag{4}$$

と変形できる．さて，特徴量マップは CNN モデルの Convolution 層によって入力画像より小さくなるため，特徴量マップを入力画像のサイズまで単純にアップサンプリングすることによって，クラス c への活性を表すヒートマップを描くことができる．この手法を用いた胸部レントゲン写真の疾病部位可視化は Wang ら [11] や Rajpurkar ら [3] によって行われている．

2.2 Grad-CAM

CAM は適用できる CNN 構造が限定的であったが，Grad-CAM[8] は最後の Convolution 層に Global Average Pooling さえ施していれば，その後の構造は自由となる．したがって，適用できる CNN が増えたため，Grad-CAM はあらゆる分野で頻繁に使用される可視化手法となった．これを可能にするため，Salvaraju らは w^c_k の定義を次のように変更した．

$$w^c_k = \frac{1}{z}\sum_x\sum_y \frac{\partial y^c}{\partial A^c_{xy}} \tag{5}$$

さらに，顕著性マップ (Saliency Map) L^c_{xy} は次のように定義される．

$$L^c_{xy} = \texttt{relu}\left(\sum_k w^c_k A^k_{xy}\right) \tag{6}$$

ここで，$\texttt{relu}(\cdot)$ は ReLU 関数で $\texttt{relu}(x) = \max(0, x)$ である．ここで得られた L^c_{xy} を入力画像までアップサンプリングして描けば，Grad-CAM によるヒートマップ可視化ができる．Grad-CAM を用いた疾病部位の可視化は最もよく行われており，Baltruschat ら [12] がその代表例である．

2.3 Grad-CAM++

Grad-CAM は式 (5) で定義されるような w^c_k を用いたことで，実は小さな特徴を無視する可能性がある．各特徴量マップ A^k_{xy} を平等に評価しているため，小さな特徴は相対的に重要度が低くなってしまう．これに対し，Grad-CAM++[9] は次のように w^c_k の定義を変えることでより良いヒートマップを描ける．

$$w^c_k = \sum_x\sum_y \alpha^{kc}_{xy}\texttt{relu}\left(\frac{\partial Y^c}{\partial A^k_{xy}}\right), \tag{7}$$

$$\alpha^{kc}_{xy} = \frac{\frac{\partial^2 Y^c}{\left(\partial A^k_{xy}\right)^2}}{2\frac{\partial^2 Y^c}{\left(\partial A^k_{xy}\right)^2} + \sum_a\sum_b A^k_{ab}\left\{\frac{\partial^3 Y^c}{\left(\partial A^k_{xy}\right)^3}\right\}}. \tag{8}$$

顕著性マップ L^c_{xy} は他と同様に $L^c_{xy} = \sum_k w^c_k A^k_{xy}$ で与えられ，入力画像のサイズまでアップサンプリングすることでヒートマップを得る．こうすることによって，全ての特徴量マップに現れる大小さまざまな特徴を平等に評価し，ヒートマップに反映することができる．これは医療画像中の小さな病変に対して他の可視化手法に比べてより大切に扱うことができることを意味する．Grad-CAM++を ChestX-ray14 に適用した例はなく，本研究の新規性の 1 つである．

2.4 Average Drop

Chattopadhyay ら [9] は，特定のデータセットに対してある可視化手法がどれほど適しているか評価する指標として Average Drop（原著では"Average Drop %"）を提案した．計算式は次である．

$$\sum_{i=1}^{N}\frac{\max\left(0, Y^{c(i)} - O^{c(i)}\right)}{Y^{c(i)}} \tag{9}$$

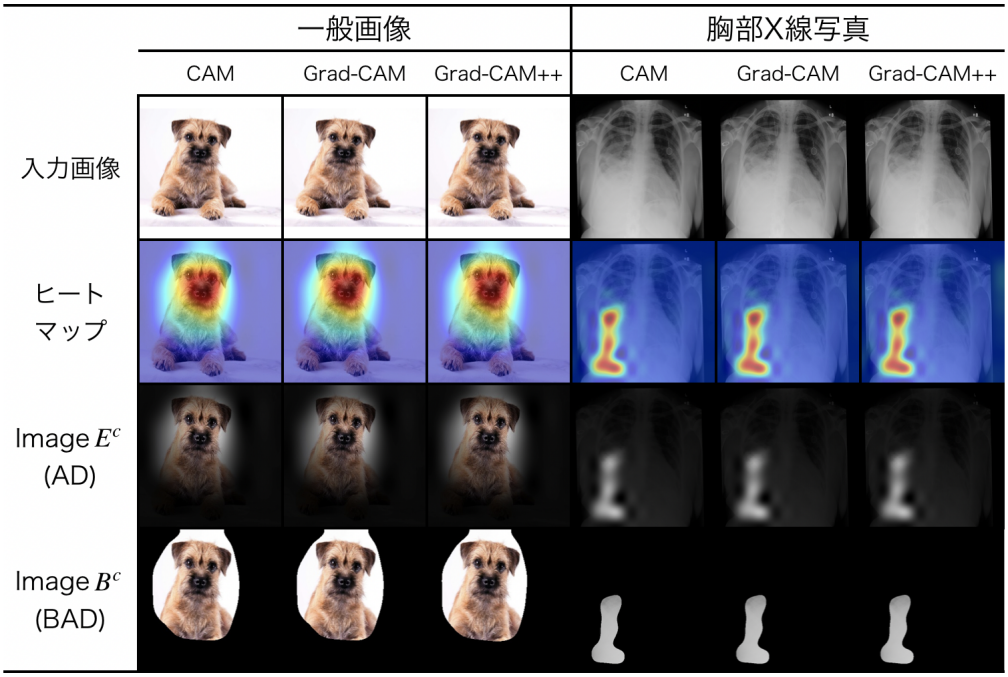

図 1 可視化結果

ここで，$Y^{c(i)}$ は i 番目の入力画像がクラス c と判断されるときの出力スコアである．また，I を入力画像とするとき，加工画像 E^c は顕著性マップを用いて $E^c = L^c \circ I$ によって作られる．この加工画像 E^c を CNN に通した時の出力スコアが $O^{c(i)}$ である．つまり，画像全体で見た時よりも，CAMs が着目した部分だけを見た時のスコアはどれ程落ちるかを測っており，これが小さいほど分類に重要な部分を可視化できていたことを意味し，良い可視化と判定される．

3. 学習結果の解釈

3.1 疾病部位の可視化

本研究では，3 つの可視化手法 CAM，Grad-CAM，Grad-CAM++について議論を行う．それぞれの手法でヒートマップを描くフローは以下の通りである．

(1) 学習データセットを用いて CNN モデルを学習し，モデルを保存する
(2) (1) のモデルを使用し，各 CAMs の顕著性マップ L^c_{xy} を計算する
(3) L^c_{xy} を入力画像のサイズまでアップサンプリングし，それを入力画像に重ねて描く

本研究ではアップサンプリングに伴う補間法は Lancozos 法を使用した

本研究では，ChestX-ray14[11] の肺炎テストデータセットと一般的な画像データセットを利用する．ChestX-ray14 は 30805 人の被験者を腹側から撮影した 112120 枚の画像を含む．なお，それぞれの画像には 14 疾病のラベルまたは 14 疾病が見つからなかったことを意味する No FIndings ラベルが付けられている．本研究では ChestX-ray14 を用いた肺炎の分類精度を競う Kaggle コンペティション [13] での分け方に従って肺炎データを学習，ヴァリデーション，テストデータの 3 つに分けた．CNN 構造については InceptionResNetV2 [14] を使用し，その Convolution 層以下を CAM を使用できるように変更した上で，Grad-CAM++を使用できるように分類層の活性化関数を ReLU または softmax にしたものとした．このモデルを学習データおよびヴァリデーションデータを使って学習，保存した後，テストデータについて可視化を行なった．

一般画像については ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) で使用された 1000 クラスの中からランダムに選んだ 100 ラベルに対し，google を用いて入手した画像 100 枚を利用する．また，Keras は ILSVRC2012 データセットを用いて学習済みの InceptionResNetV2 モデルを提供している．今回はそれを利用し，提案手法が X 線画像だけでなく一般的な画像データセットに対しても利用できることを明らかにする．

3.2 可視化結果

可視化の結果は図 1 に提示した．本研究では，InceptionResNetV2 の分類層の活性化関数が softmax であるとき Grad-CAM++を用いた可視化はできなかった．したがって以下の議論で InceptionResNetV2 の分類層の活性化関数は ReLU である[*2]．一般的な画像について，CAM と Grad-CAM は犬の鼻あたりに着目していて，Grad-CAM++は

[*2] 一般に分類タスクにおいて活性化関数を softmax ではなく ReLU にすると分類精度が低下する．今回使用したモデルは AUC が 0.93 程度のモデルである．

より広く見ていることが分かる．医療画像については，3手法間大きな違いは見て取れないが，Grad-CAM++の活性箇所が他の2手法よりわずかに大きいことがわかる．

4. 可視化手法の評価

Chattopadhyay らによる Average Drop の定義は式 (9) であるが，原著中の文脈や実験結果から正しくは次であると推測される．

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\max\left(0, Y^{c(i)} - O^{c(i)}\right)}{Y^{c(i)}} \tag{10}$$

以後，これを AD の定義として利用する．

原著 [9] では，ImageNet に含まれるような一般的な画像について利用できるとされている．しかし，AD は以下の2つの理由により医療画像には不適切である．

(1) AD は加工画像 E^c を作成する際にアルファ値を調整する
(2) 評価値は特徴の大きさによって左右される可能性がある

一般に放射線医は X 線写真の不自然な透過光の亢進，低下を発見することで診断する．よって，画像中の「白さ」「暗さ」は重要な情報であるが，普通ならば疾病が認められない部分であってもアルファ値を変更したことにより，疾病のように見えてしまう可能性がある．その場合，加工画像をモデルに通した場合に可視化手法によってではなく，アルファ値を変更したことによってスコアが低下することもあり得る．2つ目については，特徴が大きい場合は AD が作るマスクが小さくなってしまい，加工画像 E^c の入力画像と変わらなくなる．その結果，特徴の大きな画像は良いスコアとなる可能性が高い．これらの議論を踏まえ，AD を改良した新しい評価指標 Black Average Drop を提案する．

4.1 Black Average Drop

AD の弱点を解決するため，2つの改善を行う．

(1) 加工画像 E^c の定義を変更する
(2) マスクによって隠される面積を固定する

まず，L^c_{xy} を降順にソートし，全体の 上位α% となる値を閾値 η とする．これを使用し，加工画像 B^c の各ピクセル値は次のように定義する．

$$B^c_{xy} = \begin{cases} I_{xy} & (L^c_{xy} > \eta) \\ 0 & (L^c_{xy} \leq \eta) \end{cases} \tag{11}$$

ここで，α は CAM，Grad-CAM，Grad-CAM++のそれぞれに対して作ることができる加工画像 B^c_{xy} の差異が最も大きくなるように定義する．この評価指標は，あるデータセットに対して最適な可視化手法は何かを決める指標であるので，その可視化結果の差が最も分かりやすい状態で測定を行う．

表 1　CNN モデルのテスト結果

		Evaluation Score		Incorrect [%]	
	CAMs	AD	BAD	AD	BAD
ImageNet	CAM	0.1197	0.1223	5.0	8.0
	Grad-CAM	0.1197	0.1223	5.0	8.0
	Grad-CAM++	**0.1094**	**0.0932**	7.0	6.0
X-ray	CAM	**0.7400**	0.7548	100	99.4
	Grad-CAM	**0.7400**	0.7548	100	99.4
	Grad-CAM++	0.7416	**0.7416**	100	99.4

ここから先は AD と同一で，入力画像 I を CNN モデルに入れた時のスコア $Y^{c(I)}$ と加工画像 B^c_{xy} を CNN モデルに入れて得られたスコア $O^{c(i)}$ を式 (10) に入れて計算を行う．加工画像の作成方法を変化させたことで，アルファ値の変更による影響がなくなり，隠す面積も α, η によって固定したことで，面積による影響もなくすことができた．

4.2 評価結果

表 1 に AD および BAD での評価値と加工画像 E^c と B^c を用いた分類での分類ミスの割合を掲載した．ImageNet の場合も胸部レントゲン写真の場合も CAM や Grad-CAM の評価値は同一になっている．一般的な画像について，AD でも BAD でも Grad-CAM++が最善であるとの結果が出た．また, 医療画像について, 2.3 で説明した通り Grad-CAM++が最適であるべきである．AD は CAM および Grad-CAM が最適との評価をしたが BAD は Grad-CAM++が最適であるとの正しい結果を出すことができた．

5. 考察

5.1 BAD の評価について

AD も提案手法も評価値の差は大きくて 0.03 程度であった．これは加工画像 E^c や B^c について，CAMs 間の差がでにくいことに起因する．しかし，疾病部位の可視化を考えると，わずかな差が重大となることもあり得るので，この評価をすることは十分有意義である．一般画像の評価値については Chattopadhay ら [8] によって，Grad-CAM++が最適であるとの報告がされていることから，一般的な画像に対して提案手法が使用可能であることが確かめられた．一方で表 1 から本手法の限界も見受けられた．加工画像を用いた分類ミスの割合 (Incorrect[%]) については，X 線画像で非常に高い．この値が小さければ，加工画像だけでの疾病分類がうまくいった，すなわち分類において核となる部分に着目できていたことが明らかとなる．しかし，現時点では非常に高い数値となっている．評価という点においてクラス c を留めて算出するため問題はないが，この点が課題として残った．

5.2 CAMs の利用について

表 1 において，AD と BAD のいずれの評価値も CAM と

Grad-CAM の値が同一となった．また，表 1 の結果についても，CAM と Grad-CAM の可視化結果はよく似ている．これは CNN の構造に起因しており，InceptionResNetV2 の Convolution 層以下を CAM に対応できるように変更したことが理由である．これは Selvaraju らによって説明されている．概要を以下にまとめる．Grad-CAM における $\frac{\partial Y^c}{\partial A^k_{xy}}$ は次のように表現される．

$$\frac{\partial Y^c}{\partial A^k_{xy}} = \frac{1}{z} w^c_k \tag{12}$$

ここで，w^c_k は CAM における Global Average Pooling から分類層への重みで，z は特徴量マップの面積 $(x \times y)$ である．したがって 2 つの可視化手法は同一となる．一方で，Grad-CAM++は CAM や Grad-CAM によって無視された小さいかつ比較的重要な特徴を捉えたことでヒートマップの活性部位はより大きくなっていると考えられる．

5.3 softmax 版 Grad-CAM++の脆弱性

Chattopadhyay らは分類層の活性化関数を softmax にした場合の Grad-CAM++の計算方法を提案した．しかし，本実験において使用することができなかった．S^N，S^P をそれぞれ No Findings と肺炎のラベルの画像を入れたときの分類層への入力，Y^N, Y^P をそれぞれ出力スコアであり，$Y^N = \mathtt{softmax}(S^N)$，$Y^P = \mathtt{softmax}(S^P)$ に対応する．本研究での実験において，S^N は最小で約-3056，S^P は最大で約 3215 になることもあった．この場合，$\exp(S^N) = 0$ となり，$\exp(S^P) = \infty$ となり，$Y^N = 1$，$Y^P = 0$ となった．その結果，Grad-CAM++を計算する際のすべての微分値，α^{kc}_{xy}，w^c_kは 0 となる したがって，顕著性マップは $L^c_{xy} = 0$ となる．つまりヒートマップを描けていない．これは，分類層への入力値が極端に大きくなったため生じたことから，完璧な分類器である場合 softmax を活性化関数とする Grad-CAM++は使用できないことが分かった．

6. 結論

ピクセル単位でのラベルが付されていないデータセットに関する可視化手法の比較に関してあまり議論されていなかった．そこで，本研究では胸部レントゲン写真に対しても利用可能な新しい評価方法 Black Average Drop を作成し，ImageNet のような一般的な画像に対しても，ChestX-ray14 より抽出した肺炎画像に対しても利用可能であることを実験的に示すことができた．本手法が他の医療画像の疾病部位可視化や宝石分類など別分野でも活用されることを期待する．また，Grad-CAM++を ChestX-ray14 に初めて適用した．Grad-CAM++は小さな特徴を大きなものと同等の価値がある情報として扱う．小さな特徴こそ疾病の早期発見に重要な情報である医療画像の分野において，これまでの先行研究でよく使われてきた Grad-CAM は小さな特徴を無視してしまう可能性があるため，Grad-CAM++を使うべきであることも提言する．

謝辞 本研究は JST，さきがけ，JPMJPR1774，ならびに JSPS 科研費 JP19H04186 の支援を受けたものである．

参考文献

[1] 西川拓：わずか 5000 人の放射線診断専門医，見落としがん死は無くせるか，日刊工業新聞，入手先 ⟨https://newswitch.jp/p/14043⟩，(2018.8.8)．
[2] 寺崎省子：がん画像診断見落とし，新たに患者 1 人死亡　千葉大病院，朝日新聞デジタル，入手先 ⟨https://www.asahi.com/articles/ASM5Y5GN7M5YUDCB00M.html⟩，(2019.5.29)．
[3] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng.：Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning，*arXiv preprint arXiv*:1711.05225，available from ⟨http://arxiv.org./abs/1711.05225⟩, (2017)．
[4] 厚生労働省医政局事課長：人工知能（AI）を用いた診断，治療等の支援を行うプログラムの利用と医師法第 17 条の規定との関係について，入手先 ⟨https://www.mhlw.go.jp/content/10601000/000468150.pdf⟩，(2018.12.19)
[5] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller：Striving for simplicity: The all convolutional net，In *ICLR(workshop track)*，available from ⟨http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a⟩，(2015)
[6] D. Smilkov, N. Thorat, B. Kim, F. Viegas, and M.M. Wattenberg：Smoothgrad: removing noise by adding noise，In *ICML(workshop track)*，available from ⟨http://icmlviz.github.io/assets/papers/3.pdf⟩, (2017)
[7] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba：Learning deep features for discriminative localization，In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921-2929，(2016)
[8] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv 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)
[9] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian：Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks，In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*，pages 839-847，IEEE (2018)
[10] Min Lin, Qiang Chen and Shuicheng Yan：Network In Network，*arXiv in arXiv*:1312.4400，available from ⟨https://arxiv.org/abs/1312.4400⟩，(2013)
[11] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers：Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases，In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*，pages 3462-3471，IEEE (2017)
[12] Ivo M. Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp and Axel Saalbach：Comparison of Deep

Learning Approaches for Multi-Label Chest X-Ray Classification, *Scientific reports* 9(1), 6381, (2019)
[13] P. Mooney : Chest X-ray Images (Pneumonia), *Kaggle*, available from ⟨https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia⟩, Kaggle (2018)
[14] Christian Szegedy, Sergey loffe, Vincent Vanhoucke nad Alex Alemi : Incepion-v4, Inception-ResNet and the impact of Residual Connections on Learning, *arXiv preprint arXiv*:1602.07261, available from ⟨https://arxiv.org/abs/1602.07261⟩, (2016)
[15] Chollet and Fran : Keras, available from ⟨https://keras.io⟩, (2015)