2020/12/11 이유진 (youjin lee)

Paper Review

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Ramprasaath R. Selvaraju · Michael Cogswell · Abhishek Das · Ramakrishna Vedantam · Devi Parikh · Dhruv Batra

IJCV, 2019

arXiv:1610.02391v3

Abstract

우리는 많은 클래스들을 구분하는 CNN-기반의 모델들이 내리는 결정에 대해 그 모델들이 더 투명하고 설명력을 가지도록 만드는 '시각적 설명'을 생성하는 기술을 제안한다. 이 방법 - Gradient-weighted Class Activation Mapping(Grad-CAM)은 이미지 내에서 해당 개념을 예측하는데 가장 중요한 영역을 표시하는 localization map을 만들어 내기 위해서 마지막 컨볼루션 레이어로 흐르는 모든 target concept의 gradients를 사용한다.

이전의 방법들과는 달리 Grad-CAM은 어떤 구조적 변화나 재학습하는 과정 없이 CNN 기반의 모델군을 넓은 범위에 걸쳐 적용 가능하다:

(1) CNNs with fully connected layers (e.g. VGG)

(2) CNNs used for structured outputs (e.g. captioning)

(3) CNNs used in tasks with multimodal inputs(e.g. visual question answering) or reinforcement learning

우리는 high-resolution이고 class-discriminative인 시각화 구현을 위해 Grad-CAM을 미세하게 표현하는 특징을 지닌 현재 존재하는 시각화 방법들과 결합하고(Guided Grad-CAM), 이를 Image Classification, Image Captioning, Visual Question Answering(VQA) 모델들에 적용시켰다. Image Classification 측면에서 보면 이 시각화 방법들은 (a) 모델의 failure mode일 때 insight 제공 (b) weakly-supervised localization 분야에서 이전 방법들보다 뛰어난 성능을 보임 (c) adversarial한 변화에 대해 robust함 (d) 기본 모델에서 더욱 높은 신뢰성을 갖춤 (e) 데이터셋 내의 bias를 식별함으로써 모델의 generalization에 도움이 됨 의 특징을 가진다. 또한, Image Captioning과 VQA 측면에서 보면 non-attention 모델에서도 input 이미지의 구분 영역을 localize하기 위해 학습 가능하다는 것을 보여준다. 마지막으로 Grad-CAM의 설명이 사용자가 deep networks로부터 나온 예측에 대한 적절한 신뢰를 확립하도록 도움을 주는지 평가하기 위해 인간 연구를 설계하고 수행했으며, 비훈련자가 실제 둘 모두 동일한 예측을 했을 때에도 'weaker' 모델과 'strong'모델을 성공적으로 분간하는데 도움이 됨을 보여주었다.

Introduction

CNN 기반의 Deep neural network들은 computer vision 분야에서 전례없는 돌파구였지만, 개별적으로 직관적인 구성 요소로 분해할 수 없는 점때문에 그들을 해석하기 어려웠다. Interpretability matters. 현재 시스템을 신뢰하고 더 나아가 우리의 삶에 의미 있게 녹아 들도록 하기 위해서는 모델들이 예측했을 때 왜 그렇게 예측했는지에 대해 설명하는 능력을 갖는 '투명한' 모델을 만드는 것이 분명 필요하다 :

(1) AI < 사람 (e.g VQA) : [목표] failure modes를 식별하여 연구자들의 연구 방향 설정에 도움

(2) AI = 사람 : [목표] 사용자들에게 적절한 신뢰와 확신 설립

(3) AI > 사람 : [목표] 사람들이 더 나은 결정을 할 수 있도록 가르치는 기계 교육

deep residual networks(ResNet)은 200개가 넘는 레이어를 가지고 여러 도전적인 분야에서 최고의 성능을 보여주었다. 이 복잡함은 network를 해석하기 더욱 어렵게 만들었고 결과적으로 해당 network들은 Interpretability와 Accuracy 사이의 스펙트럼을 탐구하기 시작했다. Zhou는 fully-connected 레이어를 포함하지 않은 CNNs 에서만 적용이 가능한 Class Activation Mapping(CAM)을 제안했다. 이 방법은 투명성을 갖기 위해서 모델의 복잡도와 성능 간의 어느정도 트레이드 오프가 있다. 그와 대조적으로 Grad-CAM은 성능을 그대로 유지할 뿐만 아니라 모델의 구조적 변화도 필요없다는 점에서 CAM보다 더 일반화된 형태이며 광범위하게 적용될 수 있음을 시사한다.

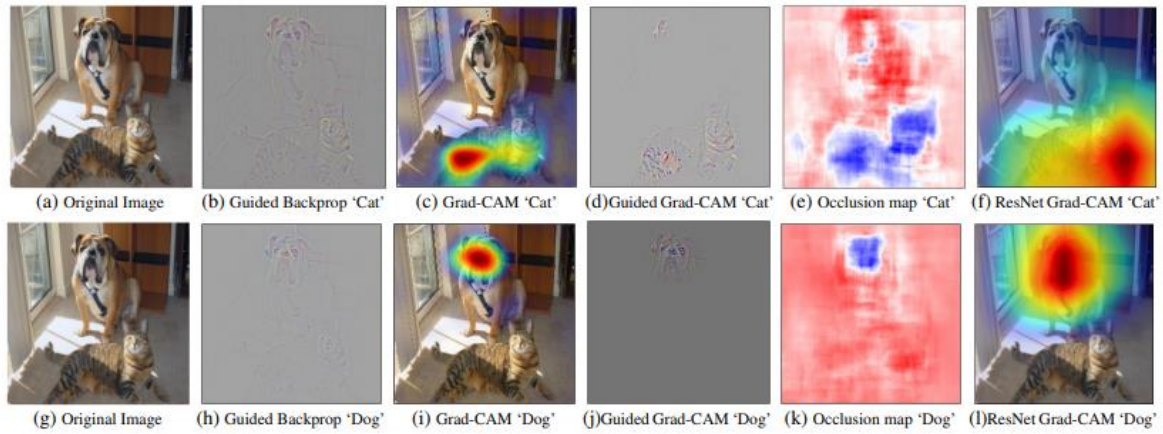


Fig. 1: (a) Original image with a cat and a dog. (b-f) Support for the cat category according to various visualizations for VGG-16 and ResNet. (b) Guided Backpropagation [53]: highlights all contributing features. (c, f) Grad-CAM (Ours): localizes class-discriminative regions, (d) Combining (b) and (c) gives Guided Grad-CAM, which gives high-resolution class-discriminative visualizations. Interestingly, the localizations achieved by our Grad-CAM technique, (c) are very similar to results from occlusion sensitivity (e), while being orders of magnitude cheaper to compute. (f, l) are Grad-CAM visualizations for ResNet-18 layer. Note that in (c, f, i, l), red regions corresponds to high score for class, while in (e, k), blue corresponds to evidence for the class. Figure best viewed in color.

"무엇이 좋은 시각적 설명을 만드는가?" 라는 질문에서 Image Classification로 생각해보면 좋은 시각적 설명은 (a) Class-discriminative이고 (b) High-resolution이여야 한다. Fig.1에서 '고양이'와 '개'에 대해 많은 시각화를 하였다. Guided Back-propagation(fig. 1b, fig. 1h)은 이미지의 디테일한 부분은 잘 살렸지만 클래스별로 구분적이지 않고, Grad-CAM(fig. 1c, fig. 1i)은 클래스별 영역이 잘 구분되어 있기는 하지만 상세하지는 않다. 그래서 우리는 Grad-CAM과 pixel-space gradient 시각화 방법들을 융합하여 (fig. 1d, fig. 1j)로 시각화하였다.

Grad-CAM

이전의 많은 연구들은 CNN의 더 깊은 표현들이 높은 수준의 시각적 구성을 포착한다고 강하게 주장해왔다. 또한, 컨볼루션 레이어들은 FC 레이어와 달리 spatial information을 보유하기 때문에 마지막 컨볼루션 레이어가 가장 많은 정보들을 가지고 있는 타협점이라고 기대한다. 해당 레이어의 뉴런들은 이미지 내의 의미적 클래스별 정보(object parts, 대상 영역)들을 찾는데, Grad-CAM은 각각 뉴런들에 중요도 점수를 할당하기 위해 CNN의 마지막 컨볼루션 레이어로 흐르는 gradient 정보를 사용한다. 우리의 방법은 deep networks의 어떤 레이어에서든 적용할 수 있지만, 이번 연구에서는 output 레이어의 결정만을 설명하는 것에 초점을 둔다.

Fig.2 에서 보이다시피, 너비 u, 높이 v인 클래스 c의 Class-discriminative인 localization map Grad-CAM $L^c_{Grad-CAM} \in \mathbb{R}^{u \times v}$을 얻기 위해서, 우리는 컨볼루션 레이어의 feature map 활성화 A^k와 관련하여 클래스 c

의 gradient 점수 y^c $\left(i.e.\ \frac{\partial y^c}{\partial A^k}\right)$를 계산한다. 이 뒤로 흐르는 gradients는 뉴런 중요도 가중치 α_k^c를 얻기 위해서 너비, 높이의 차원에 맞춰 global-average-pooling 된다.

$$\alpha_k^c = \underbrace{\frac{1}{Z}\sum_i\sum_j}_{\{global\ average\ pooling\}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\{gradients\ via\ backprop\}} \tag{1}$$

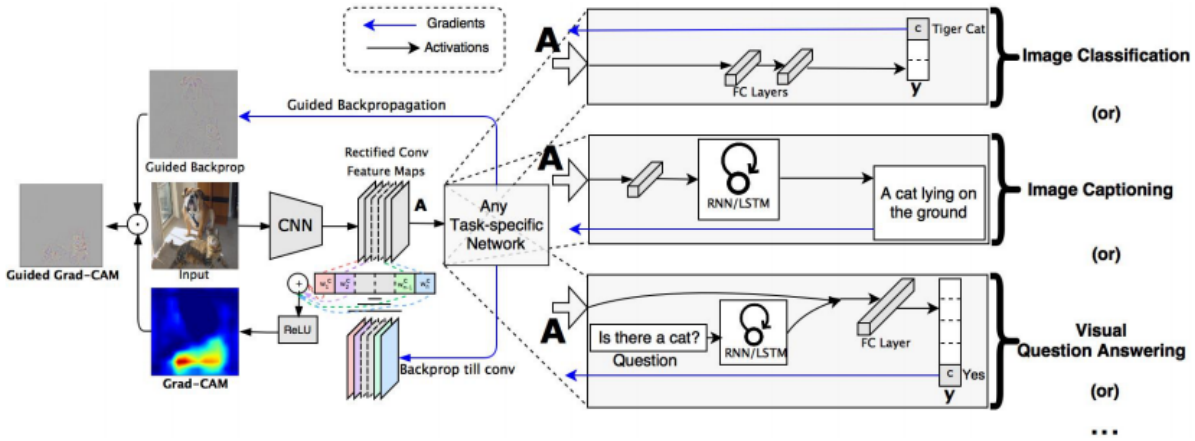


Fig. 2: Grad-CAM overview: Given an image and a class of interest (*e.g.*, 'tiger cat' or any other type of differentiable output) as input, we forward propagate the image through the CNN part of the model and then through task-specific computations to obtain a raw score for the category. The gradients are set to zero for all classes except the desired class (tiger cat), which is set to 1. This signal is then backpropagated to the rectified convolutional feature maps of interest, which we combine to compute the coarse Grad-CAM localization (blue heatmap) which represents where the model has to look to make the particular decision. Finally, we pointwise multiply the heatmap with guided backpropagation to get Guided Grad-CAM visualizations which are both high-resolution and concept-specific.

활성화에 대한 gradients를 역전파하면서 α_k^c를 계산하는 동안, 실제로 계산은 가중치 행렬과 마지막 컨볼루션 레이어까지의 활성화 함수에 관한 gradient들의 연속적인 행렬곱이 해당된다. 이런 이유로 가중치 α_k^c는 A로부터 흐르는 deep network의 부분 선형화를 나타내고, 목표 클래스 c에 대한 feature map k의 '중요성'을 포착한다. 우리는 forward 활성화 maps의 weighted combination을 수행하고, 이를 위해 RELU를 사용하였다.

$$L_{Grad-CAM}^c = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\{linear\ combination\}} \tag{2}$$

이로 인해 컨볼루션 feature map과 동일한 크기의 coarse heatmap이 생성된다. 우리는 positive 영향을 가지는 특징에만 관심이 있었기 때문에 Linear combination of maps에 ReLU를 적용했다. ReLU를 사용하지 않으면 localization map은 때때로 원하는 클래스보다 더 많은 클래스들을 강조하기 때문에 localization에서 성능이 떨어진다. 일반적으로 y^c는 이미지 분류 목적의 CNN에 의해 생성된 클래스 점수일 필요는 없다. 이는 caption의 단어나 질문에 대한 대답을 포함한 어떤 차별화된 활성화일 수도 있다.

Grad-CAM generalizes CAM

$$Y^c = \underbrace{\sum_k w_k^c}_{class\ feature\ weights} \underbrace{\frac{1}{Z}\sum_i\sum_j}_{\{global\ average\ pooling\}} \underbrace{A_{ij}^k}_{feature\ map} \tag{3 CAM}$$

F^k를 global average pooled output이라고 정의하면 $F^k = \frac{1}{Z}\sum_i\sum_j A_{ij}^k$ (4)이고, 따라서 CAM은 $Y^c = \sum_k w_k^c \cdot F^k$ (5)가 된다. (w_k^c는 c번째 클래스와 k번째 feature map을 연결하는 가중치) 우리가 얻은 feature map F^k에 클래스 c의 점수의

gradient(Y^c)를 취하면 $\frac{\partial Y^c}{\partial F^k} = \frac{\frac{\partial Y^c}{\partial A_{ij}^k}}{\frac{\partial F^k}{\partial A_{ij}^k}}$ (6) 가 된다. A_{ij}^k에 관해서 (4)의 도함수를 취하면, 우리는 $\frac{\partial F^k}{\partial A_{ij}^k} = \frac{1}{Z}$임을 알 수 있고, (6)에서 이를 대체하면 $\frac{\partial Y^c}{\partial F^k} = Z \cdot \frac{\partial Y^c}{\partial A_{ij}^k}$ (7)이 된다. (5)로부터 $\frac{\partial Y^c}{\partial F^k} = w_k^c$임을 알 수 있다. 이런 이유로 $w_k^c = Z \cdot \frac{\partial Y^c}{\partial A_{ij}^k}$ (8)이 되며, 모든 픽셀들 (i, j)에 걸쳐 양쪽 합해주면 $\sum_i \sum_j w_k^c = \sum_i \sum_j Z \cdot \frac{\partial Y^c}{\partial A_{ij}^k}$ (9)이 된다. Z와 w_k^c는 (i, j)의 영향을 받지 않기 때문에, 다시 정리하면 $Z w_k^c = Z \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k}$ (10) 이다. 여기서 Z 는 feature map의 픽셀 수이므로, 우리는 재정리하여 $w_k^c = \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k}$ (11)의 식으로 표현할 수 있다.

Guided Grad-CAM

Grad-CAM은 Class-discriminative이고 관련 이미지 영역들을 localizes하기는 하지만, Guided Backpropagation, Deconvolution과 같이 미세하게 details을 강조하는 능력은 없다. Guided Backpropagation은 ReLU를 거쳐 역전파 했을 때 negative gradient는 억제되는 이미지에 대하여 gradients를 시각화한다. 이는 뉴런을 억제하는 픽셀이 아니라 뉴런에 의해 감지된 픽셀을 포착하는 것을 목표로 한다. Fig. 1c를 보면 Grad-CAM은 쉽게 고양이를 localize하고 있지만, coarse heatmap으로는 왜 네트워크가 이 특정 영역을 '고양이'로 예측했는지는 명확하지는 않다. 양쪽 방법의 좋은 면들을 결합하기 위해서 우리는 Guided element-wise multiplication을 통해 Backpropagation과 Grad-CAM을 결합했다.(Fig. 2) Guided Backpropagation을 Deconvolution으로 대체하여 융합하는 것도 비슷한 결과를 가져오지만, Deconvolution 시각화가 artifacts를 가져온다는 점과 Guided Backpropgation이 일반적으로 덜 noisy하다는 점이 있다.

Counterfactual Explanations (Fig. 3)

(a) Original Image (b) Cat Counterfactual exp (c) Dog Counterfactual exp

Fig. 3: Counterfactual Explanations with Grad-CAM

Grad-CAM을 조금 수정하면 네트워크가 예측을 변경할 영역에 대한 뒷받침을 강조하는 설명을 얻을 수 있다. 결과적으로 이 영역에서 발생하는 개념들을 제거하면 모델이 예측에 대하여 더 자신감을 갖게 될 것이다. 우리는 이 설명법을 counterfactual explanations라 한다. 구체적으로 우리는 컨볼루션 레이어의 feature maps A에 대한 y^c의 gradient를 무효화한다. 그래서 중요도 가치 α_k^c는 아래의 식과 같이 된다.

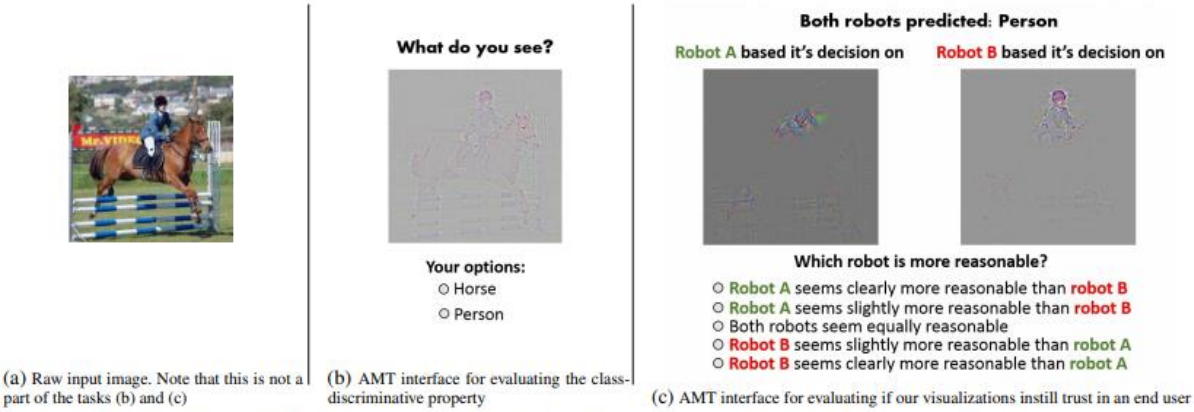
$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \underbrace{-\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{Negative gradeints}} \quad (12)$$

$global\ average\ pooling$

Evaluating Localization Visualizations

우리는 interpretability와 faithfulness의 trade-off를 이해하기 위해 인간 연구와 실험을 수행했다. 이 실험을 위해 우리는 PASCAL VOC 2007 데이터셋으로 잘 튜닝된 VGG-16와 AlexNet 모델을 비교했다. 여기서 우리는 4가지 시각화 방법 : Deconvolution, Guided Backpropagation, Grad-CAM기반의 각각 2가지 방법들 (Deconvolution Grad-CAM, Guided Grad-CAM) 으로 카테고리별 시각화를 얻었고, 이를 Amazon Mechanical Turk(AMT)에서 고용된 43명에게 "2개의 카테고리 중 이미지에 묘사된 것은 어떤 것인가?"라는 질문을 던졌다.(Fig. 5) 위 실험은 4개의 시각화로 총 90개의 이미지 카테고리 쌍을 사용하여 수행되었으며, 실

제값에 대하여 평가했고 정확도를 얻기 위해 평균값을 구했다.(Table. 2) 표를 참고하여 보면 실제로 사람들이 식별가능한 수준은 각각 44.44%, 61.23%로 16.79%나 차이가 난다.



(a) Raw input image. Note that this is not a part of the tasks (b) and (c)

(b) AMT interface for evaluating the class-discriminative property

(c) AMT interface for evaluating if our visualizations instill trust in an end user

Fig. 5: AMT interfaces for evaluating different visualizations for class discrimination (b) and trustworthiness (c). Guided Grad-CAM outperforms baseline approaches (Guided-backprop and Deconvolution) showing that our visualizations are more class-discriminative and help humans place trust in a more accurate classifier.

Method	Human Classification Accuracy	Relative Reliability	Rank Correlation w/ Occlusion
Guided Backpropagation	44.44	+1.00	0.168
Guided Grad-CAM	61.23	+1.27	0.261

Table 2: Quantitative Visualization Evaluation. Guided Grad-CAM enables humans to differentiate between visualizations of different classes (Human Classification Accuracy) and pick more reliable models (Relative Reliability). It also accurately reflects the behavior of the model (Rank Correlation w/ Occlusion).

Evaluating Localization Ability of Grad-CAM

Weakly-supervised Localization

		Classification		Localization	
		Top-1	Top-5	Top-1	Top-5
VGG-16	Backprop [51]	30.38	10.89	61.12	51.46
	c-MWP [58]	30.38	10.89	70.92	63.04
	Grad-CAM (ours)	30.38	10.89	**56.51**	46.41
	CAM [59]	33.40	12.20	57.20	**45.14**
AlexNet	c-MWP [58]	44.2	20.8	92.6	89.2
	Grad-CAM (ours)	44.2	20.8	68.3	56.6
GoogleNet	Grad-CAM (ours)	31.9	11.3	60.09	49.34
	CAM [59]	31.9	11.3	60.09	49.34

Table 1: Classification and localization error % on ILSVRC-15 val (lower is better) for VGG-16, AlexNet and GoogleNet. We see that Grad-CAM achieves superior localization errors without compromising on classification performance.

Weakly-supervised Segmentation

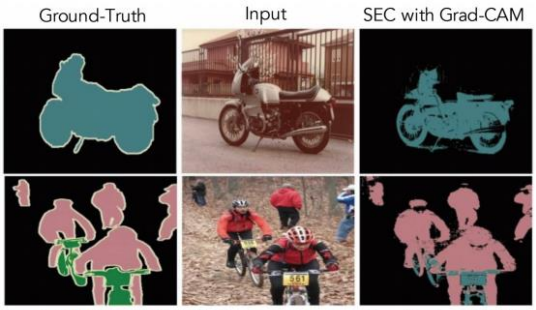


Fig. 4: PASCAL VOC 2012 Segmentation results with Grad-CAM as seed for SEC [32].

이전에는 VGG-16기반 네트워크에 CAM을 매핑했다면, 우리는 Grad-CAM으로 대체하여 적용하였다. 그 결과 Intersection over Union(IoU) score가 44.6에서 49.6으로 상승했다.

Diagnosing image classification CNNs with Grad-CAM

우리는 이미지 분류 CNNs의 failure modes에 대한 분석, adversarial noise의 영향에 대한 이해, 데이터 셋 내의 bias 식별 및 제거 목적의 Grad-CAM 사용을 입증한다.

(1) Analyzing failure modes for VGG-16 (Fig. 6)

네트워크가 잘못된 예측을 했을 때, 우리는 우선 해당 리스트를 추출하여 Guided Grad-CAM을 통해 실제 정답과 예측한 값을 시각화한다. 이를 통해 우리는 놀랍게도 전혀 합리적이지 않아 보였던 예측에 대한 합리적인 설명을 얻게 된다.

(2) Effect of adversarial noise on VGG-16 (Fig. 7)

Goodfellow는 input 이미지에 감지할 수 없는 미세한 변화를 주어 네트워크가 높은 confidence를 가지면서 잘못된 분류를 하도록 네트워크를 속이는 adversarial 예제에 대한 현재 deep networks의 취약점을 입증했다. 그러나 Fig. 7에서 보이다시피 네트워크가 'tiger cat', 'boxer'의 부재를 확신하고 있음에도 불구하고, Grad-CAM 시각화는 해당 영역을 정확하게 localize할 수 있는 것으로 보아 위 방법은 adversarial noise에 강건함을 알 수 있다.

(3) Identifying bias in dataset

우리는 "의사"와 "간호사"를 분류하는 2진 분류 VGG-16 모델을 학습시켰고, validation 정확도는 좋았지만 test 정확도는 82%로 generalization은 그렇게 잘 되어있지 않은 모델에 Grad-CAM 시각화를 수행했다. Fig. 8을 보게 되면 해당 모델은 사람의 얼굴, 머리 모양을 보고 의사인지 간호사인지 구별한 것으로 드러났는데, 이는 성 고정관념을 학습했다는 의미였다. 결과적으로 실제 학습한 데이터에서 의사 이미지의 78%가 남성, 간호사 이미지의 93%가 여성임을 알게 되었고, 해당 bias를 줄여 결과적으로 test accuracy를 90%까지 향상시킬 수 있었다.

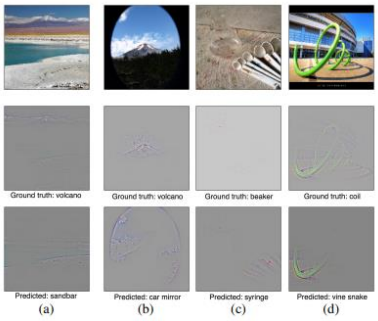


Fig. 6: In these cases the model (VGG-16) failed to predict the correct class in its top 1 (a and d) and top 5 (b and c) predictions. Humans would find it hard to explain some of these predictions without looking at the visualization for the predicted class. But with Grad-CAM, these mistakes seem justifiable.

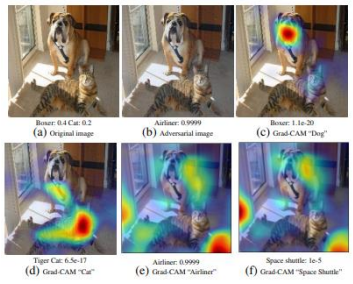


Fig. 7: (a-b) Original image and the generated adversarial image for category "airliner". (c-d) Grad-CAM visualizations for the original categories "tiger cat" and "boxer (dog)" along with their confidence. Despite the network being completely fooled into predicting the dominant category label of "airliner" with high confidence (>0.9999), Grad-CAM can localize the original categories accurately. (e-f) Grad-CAM for the top-2 predicted classes "airliner" and "space shuttle" seems to highlight the background.

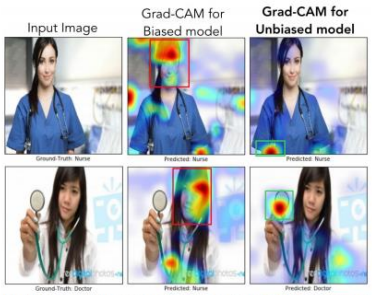


Fig. 8: In the first row, we can see that even though both models made the right decision, the biased model (model1) was looking at the face of the person to decide if the person was a nurse, whereas the unbiased model was looking at the short sleeves to make the decision. For the example image in the second row, the biased model made the wrong prediction (misclassifying a doctor as a nurse) by looking at the face and the hairstyle, whereas the unbiased model made the right prediction looking at the white coat, and the stethoscope.

Conclusion

위 논문에서 우리는 어떤 CNN 기반의 모델이든 시각화 설명을 생산함으로써 더 투명해지도록 하는 새로운 Class-discriminative인 localization 기술(Grad-CAM)을 제안한다. 또한, 우리는 고해상도이면서 Class-discriminative인 시각화를 얻기 위해 Grad-CAM localization을 현재 존재하는 고해상도 시각화 기술과 결합(Guided Grad-CAM)했다. 이 시각화 방법은 양쪽 측면(interpretability, faithfulness) 둘 다에서 현재 존재하는 방법들 중보다도 뛰어난 성능을 보인다. 인간 연구에서 우리는 우리의 시각화 방법이 클래스 간 더 정확하게 구별할 수 있음을 입증했을 뿐 아니라 데이터셋에서도 학습에 영향을 주는 bias를 식별하는데 도움이 됨을 보여주었다. 마지막으로 우리는 Grad-CAM이 image classification, image captioning, VQA 에도 적용 가능함을 보여주었다. 우리는 진짜 AI 시스템이 지능적일 뿐만 아니라 사람들이 그 시스템을 신뢰하고 사용할 수 있도록 시스템의 결정에 대해 추론할 수 있어야 한다고 생각한다.