**Introducing and assessing the explainable AI (XAI) method: SIDU**

<https://arxiv.org/pdf/2101.10710.pdf>

**0. Related Works – Visual Explanation**

|  |
| --- |
| **Saliency maps:** visual explanation을 위해 사용된다.  **Back-propagation methods:**  Feature signal을 output neuron으로부터 input neuron까지 single pass를 통해 전달한다. (Layer-wise Relevance Propagation, DeCovNet 등)  **Gradient-based methods:**  Backpropagation 알고리즘에 Gradient 또는 그것의 수정된 버전을 적용하여, CNN의 output의 detivative를 그것의 input에 대해 시각화한다. (Grad-CAM 등)   * Pixel space에서 gradient ascent를 적용하여 class-specific saliency map을 생성한다. * 이 synthesized image는 class-specific visualization으로 작용하며 주어진 CNN이 해당 class에 대해 어떻게 모델링되었는지 이해할 수 있게 한다.   **Perturbation-based methods:**  Output으로의 전체적인 변화를 트래킹하는 동안 Input이 perturb된다. (state-of-the-art RISE 등)   * 어떤 경우에는 모델의 중간 레이어에서 변화가 발생할 수 있다.   **Approximation-based method:**  Deep CNN을 보다 간단한 approximation으로 대체하며, visual explanation을 더 쉽게 생성할 수 있다. (LIME 알고리즘 등)   * Decision Tree도 같은 목적으로 이용할 수 있지만, visual input에 명시적으로 적용할 수 없다. |

**1. SIDU: Proposed Method**

|  |
| --- |
| **< Figure 2 >** |
| **NEXT PAGE** |

|  |
| --- |
| **PREVIOUS PAGE** |
| Gradient-based 모델과 perturbation explanation method에 의해 생성된 **visual explanation은 object class의 전체 salient region을 localize하는 데 실패**한다. 이를 극복하기 위해 본 논문에서는 SIDU 방법을 제안한다.   * 여기서는 **어떤 주어진 CNN 모델에 대해서도 계속적으로 더 좋은 explanation method를 제공할 수 있는 XAI method**를 제안한다. * 제안하는 방법은 **마지막 Convolution layer를 이용하여 mask**를 생성한다.   + 이들 mask로부터 Similarity Difference와 Uniqueness score가 계산되며, CNN 모델의 최종 결정에 대한 설명을 생성한다. * **< Figure 2 >** 는 SIDU에 대한 overview이다.   이 방법은 **총 3개의 step**으로 구성된다.   * 1. CNN으로부터 마지막 convolution layer를 추출하여, 그 layer를 이용하여 mask를 생성한다. * 2. 예측된 class에 대해 각 mask에 대한 Similarity Difference를 계산한다. * 3. 각 mask에 대한 weight을 계산하고, prediction을 설명할 수 있는 final map으로 합성한다. |

**2. Step 1. Generating Feature Activation Masks**

|  |
| --- |
| CNN model의 예측된 output에 대한 visual explanation을 생성하기 위해, 먼저 마지막 convolution layer로부터 feature image mask를 생성한다.   * 어떤 **deep CNN model F**에 대해서도, **크기가 n\*n\*N인 마지막 convolution layer**에 대해 다음을 수행한다. (n: conv layer의 size, N: feature의 개수)   + 각각의 feature activation map 는 binary mask 로 바뀐다.   + 이들 각각의 binary mask는 bi-linear interpolation을 이용하여 upsampling된다.   + Binary mask는 [0, 1] 사이의 값을 가지며 더 이상 binary하지 않다.   + Feature activation mask (up-sampled) 와 입력 이미지 I 간에 point-wise multiplication을 수행한다.   + Feature image mask 를 얻는다. |

**3. Step 2. Computing Similarity Differences and Uniqueness**

**4. Step 3. Explanations for the prediction**