**EXplainable AI (XAI) approach to image captioning**

[**https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9175200**](https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9175200)

|  |
| --- |
| **0. INTRODUCTION**  딥러닝의 **black-box 패러다임** 때문에 현존하는 접근 방법은 특정 이미지에 대해서 왜 특정 단어가 선택되었는지를 설명하기 어렵다. 이 한계점은 **경우에 따라 absurd(불합리한) caption을 생성**하도록 유도할 수 있다. |

|  |
| --- |
| **1. RELATED WORKS**  **Image captioning with encoder-decoder model:** 딥러닝에서 image captioning을 위해 deep **recurrent model인 encoder-decoder 모델**이 사용되었다.  **Image captioning with object detection:** 최근에는 object detection 알고리즘이 **특정 부분에 대한 더 상세한 caption**을 획득하기 위해 사용된다.  **Image captioning with attention mechanism:** **attention을 포함하는 신경망 프로세스**는 computational neuroscience에서 연구되어 왔는데, 최근에는 많은 attention 기반의 딥러닝 모델이 연구되고 있다. |

|  |
| --- |
| **2. PROPOSED MODEL - FIGURES**  모델은 **Generation part와 Explanation part**로 나뉜다.    **<Figure 2>**    **<Figure 3>** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **3. PROPOSED MODEL – GENERATION PART**  CNN 및 RNN 기반 encoder-decoder 프레임워크이다.   |  |  | | --- | --- | | Encoder | 전체 이미지에서 **feature vector를 추출**한다.   * encoder에서는 **VGG-16** 모델을 사용하여 이미지를 **fixed size**로 조절하여 feature vector를 추출한다. | | Decoder | Feature vector를 이용하여 **word를 생성**한다.   * **LSTM을 이용**하여 각 time step에서 **image feature vector와 word embedding**을 이용하여 word를 생성한다. * **Negative log likelihood loss function**을 이용하여 **image-caption pair**에 대해 encoder-decoder 모델의 학습 가능한 파라미터를 joint하게 최적화시킨다.   + 그러나 이들 pair는 **이미지의 특정 부분을 identify하지 못**하므로, **explanation part**를 통해 **important object를 탐지**할 수 있게 한다. | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **4. PROPOSED MODEL – EXPLANATION PART**  Explanation part는 **generation part가 training 또는 inferencing stage인지**에 따라 다음의 2가지 역할을 한다.   |  |  | | --- | --- | | Training stage | Image-sentence relevance loss인 **를 생성**한다.   * 이것은 생성된 **caption이 input image의 object를 얼마나 잘 고려하고 있는지**를 디지털화하여 나타낸다. * Object는 물체 탐지 알고리즘을 통해 추출된다. * **Generation part가 더 많이 학습**될수록 모델은 **object를 고려한 caption을 더 잘 생성**할 수 있다. | | Testing stage | Input image와 generation part에서 생성된 word에서 **추출된 region에 대한 weight matrix를 생성**한다.   * 각 weight value는 해당 pair에서의 **object와 word의 relevance**를 나타낸다. (아래 그림 참고) |   Explanation part는 **region-word attention 모델**과 **interpretability enhancement (IE) 모델**로 구성된다.   |  |  | | --- | --- | | Region-word attention | 물체 탐지 과정에서 생성된 region을 이용하여 **weight matrix를 생성**한다. | | Interpretability enhancement | Generation part에서 생성된 **caption이 object를 잘 나타내는지 평가**를 위하여, weight matrix를 이용하여 **image-sentence relevance loss**를 생성한다. | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **5. PROPOSED MODEL – REGION-WORD ATTENTION MODEL**  Visual attention model과 비교했을 때 이 모델은 **training procedure가 다르다**.   * 이 논문의 attention model의 목표는 **generation part가 region information을 고려하도록 돕는 것**이다. * 이것을 달성하기 위해 **attention mechanism**을 이용하며, attention model은 input region과 word에 대한 **weight matrix를 생성**한다.   **<Figure 3>** 해석   |  |  | | --- | --- | | Left side | **region과 word가 attention model에 입력**된다.   * Region은 물체 탐지 알고리즘을 통해 추출된 이미지의 특정 부분이다. * Word는 **training stage**에서 **generation part**에 의해 생성된다. | | Middle | **attention model의 구조**로, **feed-forward neural network**으로 파라미터화된다. | | Right | **weight matrix**를 나타내며, 이것은 **모델의 출력**에 해당한다.   * Weight matrix에서 **각 열은 각 region에 대한 weight vector**를 나타낸다. * 각 weight vector는 **각 region과 각 word에 대한 관련성**을 나타내며 값이 클수록 관련성이 높다. * Weight value 는 다음과 같이 계산된다.   여기서 **의 범위는** 이며 **모든 의 값의 합은 1**이다. | |

|  |
| --- |
| **6. PROPOSED MODEL – REGION-WORD ATTENTION MODEL – TRAINING PROCEDURE**  Explanation part는 **generation part의 학습 이전에 pre-train**된다.   * Pre-train을 위해서 **attention model이 먼저 훈련**되고, 그 다음에 **IE model**이 훈련된다. * Training phrase에서는 attention model의 **input**이 **이미지에서 추출된 각 region과 word**에 해당한다. * Attention model의 **output**은 **input region과 word에 대한 weight vector**이다. * Attention model의 **학습 가능한 파라미터를 최적화**하기 위하여 다음과 같은 mean square loss function을 이용한다.   여기서 similarity는 **pre-train된 word embedding**에 의해 계산되며, 값의 범위는 **-1 ~ +1**이다.   * 여기서는 -1은 2개의 word가 **embedding에서 semantically opposite**라는 것을 의미하므로, **0 to +1**의 값을 이용한다. |

|  |
| --- |
| **7. PROPOSED MODEL – INTERPRETABILITY ENHANCEMENT MODEL (IE MODEL)**  IE model은 attention model로부터 생성된 weight matrix를 이용하여 **image-sentence relevance loss인** 를 생성한다.   * IE 모델은 generation part가 caption을 생성할 때 **region information을 잘 이용하는지 평가**한다. * 그렇게 하기 위하여 모델은 **weight matrix의 각 region**에서 **가장 큰 weight value를 갖는 relevant region-word pair를 선택**한다. 이때의 region-word pair들을 이용하여 IE 모델은 각 region-word pair가 **실제로 정확**하고 **실제 데이터 분포와 관련**되어 있는지 검사한다.   + 따라서 **Bayesian추론**, IE모델 출력값은 **예측된 posterior probability** 이다.   + 이 확률은 **선택된 pair의 region과 word가 실제 데이터 분포와 관련되는지**를 나타내므로, **가 주어진 와 에 대하여 크다**는 것은 **가 실제 데이터 분포에서의 과 유사성이 크다**는 것을 의미한다.   + Conditional distribution을 알지 못하기 때문에 posterior probability는 직접 계산할 수 없으므로, **prior probability에 기반한 Bayesian 추론을 통해 그 값을 추론**한다.   + 따라서 는 다음과 같이 계산된다. |

**(NEXT PAGE)**

**(PREV PAGE)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **7. PROPOSED MODEL – INTERPRETABILITY ENHANCEMENT MODEL (IE MODEL)**  그런데 **와 는 벡터 형태**이므로 모든 확률 값을 직접 구할 수 없으며, 따라서 **region과 word에 대한 vector를 입력**으로 하는 다음과 같은 모델을 만든다.    **<모델 설명>**   * **Feed-forward neural network로 파라미터화**되어 있다. * Weight matrix에서 **region-word pair**들이 선택되고, 이것들은 **IE model**에서 사용된다. * IE 모델을 훈련시키기 위하여 **cross-entropy loss function**을 사용한다.  |  |  | | --- | --- | | Truth value | **Likelihood**와 **prior probability**의 곱 | | Predicted value | IE 모델로부터 생성된 **posterior probability** | |