JESK J Ergon Soc Korea 2023; 42(6): 661-673
https://dx.doi.org/10.5143/JESK.2023.42.6.661
https://jesk.or.kr eISSN:2093-8462

A Quantitative Analysis for Effects of Data Biases on AI Model Outcomes

Inheon Choi, Sangwon Lee

Korea University, Department of Industrial and Management Engineering, Seoul, 02841

데이터의 편향성이 인공지능 모델 결과에 미치는 영향에 대한 정량적 분석

최인헌, 이상원

고려대학교 산업경영공학과

Corresponding Author
Sangwon Lee
Korea University, Department of Industrial and Management Engineering, Seoul, 02841
Email : upcircle@korea.ac.kr

Received : October 31, 2023
Revised : November 06, 2023
Accepted : November 07, 2023

Objective: The objective of the study is to investigate how biases in emotion data affect the inference process of artificial intelligence (AI) models. The first study visually examines the inference process of two distinct models trained on positively biased and negatively biased data, respectively, to observe differences. The second study investigates captions generated by image captioning models trained on four datasets with different ratios of positive and negative biases in emotion.

Background: As AI development accelerates with diverse and large-scale data, concerns about biases raising erroneous learning and reduced prediction accuracy have emerged. Data biases cause unfair outcomes related to race, gender, or ideology.

Method: The first study explores the influence of emotion data bias on a CNN model's inference process, revealing concentration patterns for positively and negatively biased emotion data using LayerCAM. The second quantitatively analyzes bias influence on outcomes using an image captioning model trained on differently biased datasets. We quantitatively analyze the relationship between the generated captions and the datasets in terms of the Sentiment Intensity Analysis (SIA) values.

Results: Divergent inference processes were observed in the first study, showing that the focused regions of the models were different by different biases in emotion data trained. The second study demonstrated that as the bias ratio shifted, the model-generated captions changed accordingly, with a tendency to produce more positive captions with decreased positive and increased negative biases.

Conclusion: This study underscores the critical impact of emotion data bias on AI model inference and outcomes. Recognizing and mitigating biases in emotion data is imperative for developing fair and effective AI models, particularly in applications involving human-AI interaction. In the context of AI integration in services, addressing emotion data bias is crucial for responsible AI deployment, especially in human-AI interaction.

Keywords: Emotion data, Data bias, AI model inference, Human-AI interaction

Copyright@2023 by Ergonomics Society of Korea. All right reserved.

1. Introduction

인공지능을 학습시키기 위한 데이터가 다양해지고 그 규모가 커짐에 따라 인공지능의 발전 또한 가속화되고 있다. 기존의 수리적인 알고리즘으로 답을 도출하기 힘든 문제들을 해결하기 위해 다양한 인공지능 모델이 연구됐고 이를 학습시키기 위한 데이터 또한 구체화되고 방대해져 왔다(Muthukrishnan et al., 2020). 하지만 이러한 발전 과정에서, 데이터의 불균형 혹은 편향으로 인해 잘못된 학습이 이루어지는 문제점이 발생하였고 인공지능 모델의 예측 정확도를 감소시켰다(Santos et al., 2018). 예측 정확도에 미치는 영향 외에도, 특정 인종, 성별 혹은 사상에 치우친 결과를 도출하는 경우도 빈번하게 발생하였다. 이러한 이유로 인하여 인공지능을 활용한 서비스나 제품을 설계함에 학습 데이터의 불균형은 중요하게 고려되어야 할 요소가 되었다. 특히 인공지능을 융합한 서비스와 기술의 개발이 활발하게 이루어지고 있는 시점에서 사람과 교감하는 인공지능이 중요하게 고려되고 있다. 인공지능에 학습된 감정을 주입함으로써 사용자가 더욱 자연스럽게 느끼게 하고자 하는 노력을 시중의 다양한 서비스를 통해 확인할 수 있다. 즉, 감정이라는 요소는 사람과 사람 뿐만 아니라 사람과 인공지능의 상호작용에도 큰 영향을 미치고 있으며 이에 대한 불균형이 발생하였을 때 인공지능 모델에 어떤 변화가 일어나는지 살펴보아야 한다. 이러한 점에 착안하여 본 연구에서는 감정 데이터의 편향성이 인공지능 모델의 추론 과정에 어떻게 영향을 미치며, 그 결과가 데이터의 편향에 어떻게 영향을 받는지 두 가지 연구를 통해 살펴보고자 한다.

첫 번째 연구는 인공지능 모델의 추론 과정에 데이터의 편향이 어떤 영향을 미치는지 알아본다. 동일한 구조의 CNN (Convolutional Neural Network) 모델이 긍정과 중립 감정을 가지는 데이터만을 통해 학습하였을 때와 부정과 중립 감정을 가지는 데이터를 통해 학습시키게 된다. 서로 다른 감정의 데이터를 학습한 모델이 결과를 도출하는 과정에 차이가 있음을 보이기 위해 설명 가능한 인공지능 기법의 하나인 GradCAM을 활용한다. 해당 연구를 통해 두 CNN 모델이 결과를 추론하는 과정에서 모델이 주로 집중하는 지역적 특성이 서로 다름을 확인하였다. 이를 통해, 데이터의 편향이 긍정과 부정이라는 특징으로 나뉘었을 때, 모델의 추론 과정이 전혀 다름을 알 수 있었다.

두 번째 연구는 첫 번째 연구에서 발견한 사실을 기반으로 감정 데이터가 편향된 정도에 따라 결과가 어떻게 영향을 받는지 정량적으로 살펴본다. 사람의 표정 데이터를 이용하여 비교적 특정한 데이터를 이용하여 학습을 진행하였던 첫 번째 연구와는 다르게, 두 번째 연구에서는 풍경, 사물 등의 비교적 일반적인 이미지와 그에 따른 캡션(caption)을 레이블(label)로 활용하는 데이터 통해 이미지 캡셔닝(image captioning) 모델을 학습시킨다. 첫 번째 연구에서 데이터를 긍정과 부정만으로 데이터를 분류하여 학습시킨 것과는 다르게, 두 번째 연구에서는 데이터의 편향을 비율적으로 서로 다르게 구성한 부분 집합들을 모델 학습에 사용하였다. 학습 결과를 통해 각각의 모델들이 생성한 캡션들이 감정 데이터의 편향 비율에 따라 차이가 발생한 것을 확인하였다. 테스트 데이터의 이미지들을 이용하여 생성된 캡션들을 분석한 결과, 학습 데이터의 긍정에 대한 비율이 감소하고 부정에 대한 비율이 증가할수록 모델이 캡션을 생성할 때 긍정적으로 생성하는 경향이 있음을 확인하였다.

2. Related Work

2.1 Emotional AI

사람의 감정은 언어적 측면에 있어서 가장 풍부한 차원을 가지는 요소이다(Mathews et al., 2016). 이 때문에 많은 심리학자들은 감정이라는 복잡한 차원적 요소를 최대한 간단하게 표현하고자 하였고, 다양한 방법들이 연구됐다(Russell, 1991; Fontaine et al., 2007). 그 중 가장 범용적으로 쓰이는 모델은 2-Dimensional Arousal-Valence Plane으로서, 긍정적인 감정과 부정적인 감정을 2차원 좌표 평면에 표시하게 된다.

이렇듯, 과거로부터 감정이라는 영역에 관해 많은 연구가 이루어져 왔다. 감정을 활용한 인공지능 서비스들이 활발하게 출시되고 있는 상황에서 인공지능에서 감정이 어떻게 다루어져야 하는지 살펴보았다. Peter and Ho (2022)에 따르면, 사회에 감정이 있는 인공지능을 도입하면서 5가지 주요한 논의 점을 제시하였다. 비윤리적이거나 악의적인 오용으로 이어질 수 있는 데이터 추적 기술에 대한 문제가 존재하고, 이러한 감정 인공지능 기술이 국가와 문화의 경계를 넘어 문화적 긴장이 발생할 수 있으며, 이에 대한 산업계의 표준이 부족하다고 주장하였다. 또한, 감정 인공지능에 대한 기존의 윤리적인 프레임워크는 종종 모호하다는 문제점이 있으며, 감정 인식 분야에서의 불완전한 완성으로 인하여 아직도 사람의 감정이 형성되는 과정조차 명확하게 설명하지 못하기 때문에 감정이 담

긴 인공지능을 사회에 적용하기는 이르다는 주장을 하였다(Barrett, 2017). 과거 연구들부터 최근 연구에 이르기까지 감정을 가진 인공지능과 사람 사이의 상호작용이나 감정을 지닌 인공지능에 대한 사람들의 반응에 대한 실험적 연구들이 활발하게 이루어져 왔다(Vallverdu et al., 2016; Hortensius et al., 2018). 하지만 실험적 연구들은 실제 모델이 감정이라는 기준에 따라 어떻게 변화하는지 알아보는 데에 한계가 존재한다. 실제 감정을 학습하는 인공지능 모델들의 경우도 감정 요소를 포함하는 데이터를 이용하여 모델을 학습시키고 감정을 포함하는 결과를 나타내는지 확인하는 수준의 연구들이 주를 이루었다. 따라서 감정이 담긴 인공지능에 대한 관찰과 실험이 필요하다는 부분에서 본 연구의 필요성이 더욱 강조되었다.

2.2 Biased AI models with imbalanced data

감정 분석 혹은 감성 분석을 다루는 인공지능 모델들은 주로 자연어 처리 모델로부터 출발하였다(Chakriswaran et al., 2019). 모델을 학습시키기 위한 데이터를 만드는 과정에서 수많은 말뭉치에 감정 레이블을 레이블링하기 시작했고 이는 데이터 수집 단계에서부터 전처리 과정에 이르기까지 다양한 모델 학습 준비 단계부터 감정을 추출하기 위한 노력이 있었다(Al-Moslmi et al., 2017).

Sharma et al. (2020)의 연구에 따르면, 최근 이미지의 특징을 추출하는 CNN과 자연어 처리에 많이 쓰이는 RNN 혹은 LSTM을 결합이 있어 이미지에 대한 캡션을 생성하는 모델들에 관한 연구가 활발하게 이루어지고 있다. 주로 이미지에 대한 설명문을 캡션으로 생성하여 물체 인식(object detection)과 이를 통한 RNN에서의 캡션 생성 모델에 대한 연구가 이루어졌다(Vinyals et al., 2015). 또한, 감정이 있는 이미지 캡션 데이터를 활용하여 이미지 캡셔닝 과정을 단순한 설명문을 만들어내는 것이 아닌 감정에 대한 요소를 가지고 있는 캡션을 생성하는 모델 또한 연구되고 있다. 하지만 해당 연구들은 긍정적인 캡션과 부정적인 캡션에 대해 이분법적으로 접근하였다는 문제점을 가지고 있다(You et al., 2018). 본 연구에서는 위와 같은 문제점에 착안하여 긍정적인 데이터와 부정적인 데이터를 일정한 비율로 분리하여 모델의 학습을 진행하고 학습 결과에 대한 분석을 진행하고자 한다.

2.3 CNN visualization

딥러닝 모델들은 일부 기계 학습 모델들과 다르게 분류, 예측 결과에 대한 원인 분석이 힘든 블랙박스 구조이다. 이로 인해 인공지능이 결과를 도출하는 과정에 대해 투명하게 알 수 있도록 하는 알고리즘에 관한 연구들이 이루어져 왔다. 이를 설명 가능한 인공지능(explainable artificial intelligence)이라는 하나의 연구 분야로서 여겨지며 다양한 딥러닝 모델의 결과 도출 과정을 투명하게 보고자 하는 노력이 이어지고 있다. 그중 CNN 모델을 중심으로 CNN 시각화(CNN visualization) 알고리즘을 본 연구에 사용하고자 한다. CNN 시각화는 폐쇄적인 CNN 모델의 추론 과정 혹은 원인에 대해 이미지에 히트맵으로 표시해줌으로써 시각적으로 추론 과정과 원인에 관해 확인할 수 있는 알고리즘이다. 앞서 모델을 학습할 때 긍정과 부정이라는 서로 다른 편향성을 가지는 데이터를 바탕으로 감정이라는 레이블에 대해 편향적으로 학습하게끔 설계하였다. 이때 두 모델이 각각 같은 레이블을 가지는 데이터에 대해 분류할 때 추론 과정에 있어서 명확한 차이를 보인다면 데이터의 편향성에 의해 모델 또한 편향될 수 있다고 시각적으로 확인할 수 있으므로 CNN 시각화 알고리즘을 본 연구에 활용하였다.

CAM (class activation mapping)은 CNN 모델의 데이터 추론 과정을 히트맵 형태로 나타내는 시각화 알고리즘이다(Zhou et al., 2016). CNN 시각화 알고리즘의 기반이 되는 이론이지만 CNN 구조의 가장 마지막 층이 fully-connected 구조인 경우에만 적용할 수 있다는 단점이 있다. 이를 보완하기 위해 역전파 그래디언트를 이용한 GradCAM 알고리즘이 연구되었지만 CNN의 마지막 층에서만 정보를 추출하기 때문에 결과에 가까운 층일수록 앞선 층들에 비해 보다 넓은 관점에서의 특징을 추출하는 CNN 모델의 특성상 자세한 시각화가 힘들다는 단점이 존재한다(Selvaraju et al., 2017).

위의 문제점들을 보완하여 연구된 알고리즘이 LayerCAM이다. LayerCAM은 여러 층으로 이루어진 CNN 모델에서 모든 층 혹은 원하는 층의 추론 정보를 시각화할 수 있다. 이러한 특징으로 인해 GradCAM 보다 자세한 시각화가 가능한 장점이 있는 알고리즘이다. LayerCAM의 특징을 이용하여 서로 다른 모델이 이미지를 분류할 때 이미지의 어떤 지역적 특징에 집중하는지 자세한 시각화를 할 수 있다. 이를 통해 각 모델의 차이를 살펴보고 모델 편향성에 대한 해석이 가능해진다(Jiang et al., 2021). CNN 시각화 알고리즘은 첫 번째 연구에서 감정을 기준으로 서로 다르게 편향되어 있는 모델이 결과를 도출하는 과정에 대해 살펴보기 위하여 사용하고자 한다.

2.4 Image captioning

이미지 캡셔닝 모델 (image captioning model)은 이미지의 특징을 추출하는 CNN 모델의 구조에 자연어 처리에 특화된 RNN 모델의 구조를 이어 이미지로부터 캡션을 생성한다. 자연어 처리 중에서도 번역에 특화되어 있는 RNN은 인코더(encoder)와 디코더(decoder)로 나누어져 있다. RNN의 인코더는 번역하고자 하는 문장을 입력 받아 특징을 추출하여 디코더로 정보를 전달하게 된다. 입력에 대한 정보를 전달받은 디코더는 해당 정보를 일종의 해독 과정을 통하여 번역 결과를 출력하게 된다(Cagrandi et al., 2021).

CNN과 RNN을 결합한 이미지 캡셔닝 모델은 위의 설명에서 RNN의 인코더 부분을 CNN으로 대체하여 자연어 처리에서 번역하고자 하는 문장에 대한 특징이 아닌 캡션을 생성하고자 하는 이미지에 대한 특징들을 RNN의 디코더로 전달하게 된다. 결과적으로 RNN의 디코더는 이미지의 특징들에 대한 캡션을 생성하게 된다. 본 연구에서는 CNN 모델로 InceptionV3를, RNN 모델로 Bahdanau 어텐션(attention) 메커니즘을 사용하였다(Szegedy et al., 2016; Bahdanau et al., 2014).

이미지 캡셔닝 모델 학습을 위하여 파이썬 신경망 라이브러리인 Tensorflow의 사전 학습된 InceptionV3 모델을 활용하였다. InceptionV3 모델은 CNN의 일종으로, 기존의 CNN 모델들보다 적은 계산량으로 효율적인 학습이 가능한 모델이다. 계산량이 적고 모델의 크기가 기존의 모델과 비교하여 컴퓨팅 자원 소비가 적다는 장점이 있다. 학습의 정확도를 보장하며 정밀한 학습보다는 용이한 학습이 결과에 대한 분석에 집중한 본 연구에 적합하여서 두 번째 연구의 특징 추출 과정에 활용하였다.

위의 문제점들을 보완하여 연구된 알고리즘이 LayerCAM이다. LayerCAM은 여러 층으로 이루어진 CNN 모델에서 모든 층 혹은 원하는 층의 추론 정보를 시각화할 수 있다. 이러한 특징으로 인해 GradCAM 보다 자세한 시각화가 가능한 장점이 있는 알고리즘이다. LayerCAM의 특징을 이용하여 서로 다른 모델이 이미지를 분류할 때 이미지의 어떤 지역적 특징에 집중하는지 자세한 시각화를 할 수 있다. 이를 통해 각 모델의 차이를 살펴보고 모델 편향성에 대한 해석이 가능해진다(Jiang et al., 2021). CNN 시각화 알고리즘은 첫 번째 연구에서 감정을 기준으로 서로 다르게 편향되어 있는 모델이 결과를 도출하는 과정에 대해 살펴보기 위하여 사용하고자 한다.

3. Study 1: CNN Models with Emotionally Biased – Binary Classification

3.1 Emotional image dataset

본 연구에서 CNN에 대한 실험을 진행하기 위하여 학습에 쓰일 수 있는 데이터에 대한 검토가 이루어졌다. 검토 기준으로서 데이터의 편향성을 명확하게 고려할 수 있는 요소가 있어야 하며, 이미지를 구성하는 지역적 특성에 공통점이 존재해야 함을 고려하였다. 해당 기준들을 바탕으로 사람의 표정 이미지 데이터를 검토하였다. 표정 이미지는 데이터들이 눈, 코, 입 등의 유사한 지역적 특성이 있어서 모델 간의 차이를 살펴보기 쉽다. 거기에 더해, 앞으로의 인공지능 서비스에 중요하게 고려되어야 할 감정에 대한 레이블이 적용되어 있어서 편향성을 명확하게 고려할 수 있는 요소가 있어야 하는 기준과 앞의 기준을 동시에 만족하는 AffectNet 이미지 데이터를 CNN을 학습시키기 위한 데이터로 선정하였다(Mollahosseini et al., 2017). AffectNet은 총 8개의 감정 레이블(happy, surprise, sad, fear, disgust, anger, contempt, neutral)로 구성된 표정 이미지 데이터이다.

첫 번째 연구의 목적은 이미지 감정 데이터를 긍정 또는 부정이라는 편향성을 가진 두 집합의 데이터로 만들고 각각의 집합을 같은 구조의 CNN 모델이 학습했을 때 어떤 구조적인 차이를 보이는지 살펴보는 데에 있다. 이를 위해 8개의 감정 레이블을 긍정, 중립, 부정의 3가지 레이블로 재구성하였다. 레이블의 재구성을 위한 객관적인 지표로서 valence-arousal 모델을 활용하였다. 해당 모델은 감정의 특징들을 축으로 정의하고 그 특징 값으로 어떤 감정을 표현하는 방법이다. 2개의 축 중 valence 축은 감정의 긍정 혹은 부정적인 정도를 나타내는데 예를 들어, 공포의 경우 아주 부정적인 valence를, 지루함이나 흥분감은 중간 정도의 valence를, 행복이나 편안함은 긍정적인 valence를 가진다(Mehrabian and Russell, 1974). 이 모델을 참고하여 8개의 레이블 중 happy, surprise 레이블을 긍정 레이블로, sad, fear, disgust, anger, contempt 레이블을 부정 레이블로, neutral 레이블을 중립 레이블로 재구성하였다.

데이터 재구성 후, Table 1과 같이 실험을 위해 데이터를 다시 2개의 부분 집합으로 분리하였다. 긍정 부분 집합은 임의로 추출한 긍

정 레이블과 중립 레이블 각 3만 장씩 총 6만 장의 이미지로 구성하였고 부정 부분 집합은 긍정 부분 집합과 마찬가지로 임의 추출한 부정 레이블, 중립 레이블 각 3만 장씩 총 6만 장의 이미지로 구성하였다.

Table 1. Separated image dataset

	Positive	Neutral	Negative	Total
Positive image dataset	30,000	30,000	0	60,000
Negative image dataset	0	30,000	30,000	60,000

3.2 CNN model structure

이미지 데이터를 학습하는 데에 효과적이고 시각화 알고리즘을 통해 서로 다른 모델의 차이점을 확인할 수 있어야 하므로 CNN 기반의 알고리즘을 통해 실험을 진행하였다. CNN은 주로 이미지 데이터를 학습하여 다양한 결과를 도출하는 데에 사용된다(LeCun et al., 1998). 학습하는 과정에서 이미지의 지역적 특징 추출(local feature extraction)을 통해 특징 지도(feature map)을 생성하고 이를 바탕으로 이미지를 판단하게 된다. 이러한 이유로 인하여 이미지를 판단하는 모델의 경우 많은 연산량이 필요하고 정확도 향상을 위해서 깊은 depth, 많은 필터 개수, 높은 해상도의 이미지 데이터가 필수적인 요소이다. 위의 3가지 요소들의 최적 조합을 AutoML을 통해 찾고 적은 연산량으로 기존 CNN 모델들과 비교하여 성능적으로 우위에 있는 EfficientNet을 이용하여 각 부분 집합을 학습하였다(Tan and Le, 2019). 모델의 이름에서 유추할 수 있듯이 다른 CNN 모델들보다 학습에 필요한 자원적 효율성이 뛰어난 모델이다. 기존 모델들보다 약 6.1배 빠르고 8.4배 적은 용량을 가지고 더 높은 정확도를 가진다. 보다 효율적인 학습을 위해 본 연구에는 EfficientNet을 활용하였다(Figure 1).

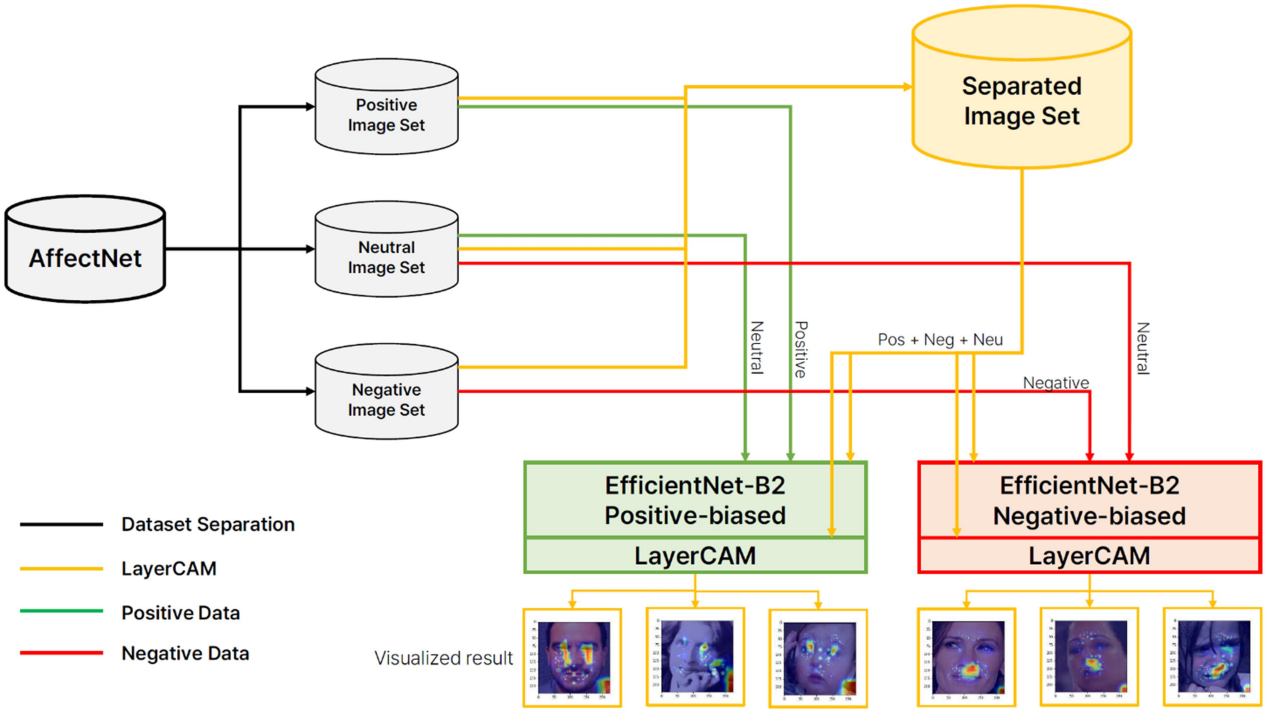


Figure 1. Overall structure of study 1

3.3 Results

긍정에 대한 데이터를 학습한 모델(P)과 부정에 대한 데이터를 학습한 모델(N)이 결과를 예측하는 과정에서 이미지의 어떤 지역 즉, 얼굴 부위 중 어디에 집중하여 결과를 도출하였는지 살펴보기 위해 학습이 완료된 두 모델들이 임의의 이미지에 대해 레이블을 예측하도록 했다. Figure 2와 Figure 3은 각각은 긍정에 대한 데이터를 학습한 모델인 P의 LayerCAM 결과와 부정에 대한 데이터를 학습한 모델인 N의 LayerCAM 결과이다.

모델 P와 모델 N은 각각 긍정으로 편향된 데이터(부정 데이터가 포함되지 않은 데이터)와 부정으로 편향된 데이터(긍정 데이터가 포함되지 않은 데이터)를 이용하여 학습이 완료된 CNN 모델이다. 같은 구조를 가진 CNN 모델이 감정을 기준으로 편향된 데이터를 통해 학습하였을 때 명확한 차이가 발생함을 Figure 2와 Figure 3를 통해 확인할 수 있다. 모델 P의 경우, 이미지를 분류할 때 얼굴 부위 중 양안이 판단 과정에 가장 큰 영향을 미쳤음을 알 수 있고 모델 N의 경우, 이미지를 분류하는 과정에 얼굴의 입 부분이 가장 큰 영향을 미쳤다. 특히, Figure 2와 Figure 3의 첫 번째 사진은 동일한 사진을 P와 N이 분류할 때 이미지 데이터의 서로 다른 지역적 특성이 영향을 미침을 명확하게 확인할 수 있다.

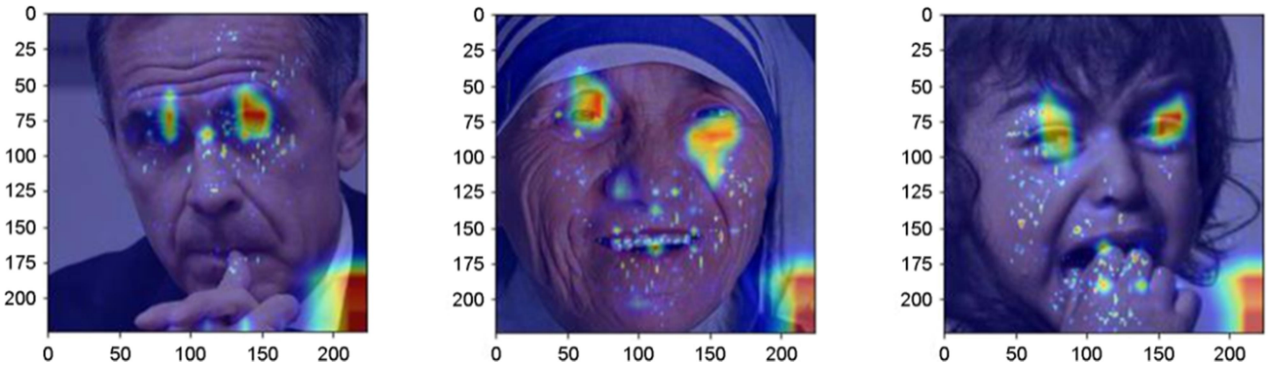


Figure 2. LayerCAM results of positive model

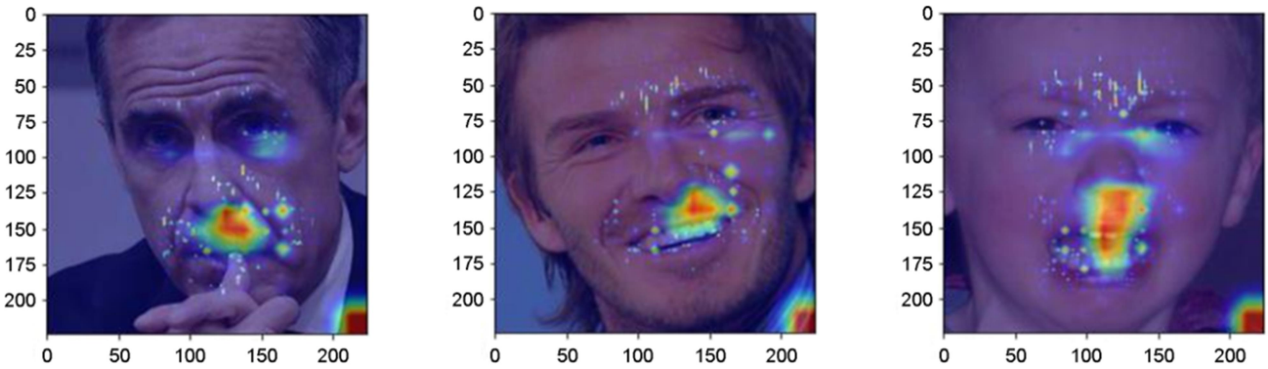


Figure 3. LayerCAM results of negative model

위의 결과들을 바탕으로, 학습 과정에서 이미지 데이터의 편향이 감정이라는 요소를 기준으로 이루어졌을 때 CNN 모델의 추론 과정에 영향을 미치는 것을 알 수 있다. 이는 데이터의 편향이 인공지능 모델의 추론 과정에 영향을 미쳐 모델의 구조적인 차이를 만들어 냄을 알 수 있다.

4. Study 2: Image Captioning Models with Emotionally Biased – Natural Language Generation

4.1 Image caption datasets

(a) (b) (c)

Figure 4. SentiCap and MS-COCO images

본 연구에서 이미지 캡셔닝에 대한 실험을 진행하기 위하여 우선적으로 학습에 쓰일 수 있는 충분한 양의 이미지 캡션 데이터와 감정 캡션이 레이블링 되어있는 이미지 데이터를 검토하였다. MS-COCO 데이터는 20만개 이상의 이미지와 그에 따른 캡션들로 이루어진 데이터로서, 물체 인식과 이미지 캡셔닝 학습을 위해 만들어진 대규모 이미지 데이터이다(Lin et al., 2014). 기존의 PASCAL과 ImageNet과 같은 이미지 캡션 데이터와 비교하여 충분한 양의 데이터를 확보할 수 있다(Everingham et al., 2015; Deng et al., 2009).

MS-COCO 데이터는 감정이 담겨있지 않고, 이미지의 물체와 배경에 대한 설명들로 캡션이 구성되어 있다. Mathews et al. (2016)의 연구는 MS-COCO 데이터의 일부를 활용하여 감정이 포함된 이미지-캡션 데이터를 생성하였다. MS-COCO 데이터의 일부를 추출하여 Amazon Mechanical Turk의 설문자들을 통하여 기존의 데이터를 구성한 캡션의 약 2,000개의 문장을 감정을 표현하는 캡션들로 치환하여 SentiCap 데이터를 구성하였다. 해당 연구에 따르면, 기존에 단순한 설명문으로 구성되어 있던 MS-COCO만을 사용하여 학습한 모델들은 감정이 없는 설명문에 가까운 캡션들을 생성하는 반면, SentiCap을 사용하여 학습한 모델들은 88.4%에 해당하는 캡션들이 감정을 가진 캡션들임을 확인하였다. 해당 연구를 근거로 본 연구에서는 MS-COCO 데이터를 기반으로 하는 SentiCap 데이터를 활용하여 학습을 진행하였다.

이미지 캡셔닝에 있어서 MS-COCO 데이터는 가장 범용적으로 쓰일 수 있는 양질의 데이터이다. 그러나 이미지들에 대한 캡션들이 감정을 포함하고 있지 않아 순수하게 MS-COCO 데이터만을 활용하여 본 연구에 사용하기에는 적합하지 않다. MS-COCO 데이터를 기반으로 일부의 데이터를 감정이 포함된 캡션으로 대체된 SentiCap을 MS-COCO와 함께 활용하여 연구를 진행한다.

Table 2. Separated image caption dataset

	SentiCap	P8N2	P6N4	P4N6	P2N8
# of positive captioned images	4,892	4,892	4,892	2,651	994
# of negative captioned images	3,977	1,223	3,261	3,977	3,977
# of neutral captioned images	0	11,133	11,133	11,133	11,133
# of total captioned images	8,869	17,248	19,286	17,761	16,104

CNN의 특징 추출과 RNN 학습을 위하여 MS-COCO 데이터와 SentiCap 데이터를 활용하여 데이터를 새롭게 구성하였다. 기존 감정 이미지 캡셔닝 연구에서의 문제점으로 제시한 데이터의 지나친 편향을 피하고 다양한 조합에 대한 학습을 진행하기 위하여 SentiCap 데이터를 총 4개의 부분 집합으로 나누었다. Table 2와 같이 긍정 캡션과 부정 캡션의 비율을 각각 8:2 (positive 80%, negative 20%; P8N2), 6:4 (positive 60%, negative 40%; P6N4), 4:6 (positive 40%, negative 60%; P4N6), 2:8 (positive 20%, negative 80%; P2N8)로 나누어 각 집단이 긍정과 부정에 대한 데이터를 모두 가지며 편향성을 띠도록 설계하고 긍정적 캡션을 1로, 부정적 캡션을 -1로, 중립적 캡션을 0으로 레이블링하였다.

4.2 Image captioning model structure

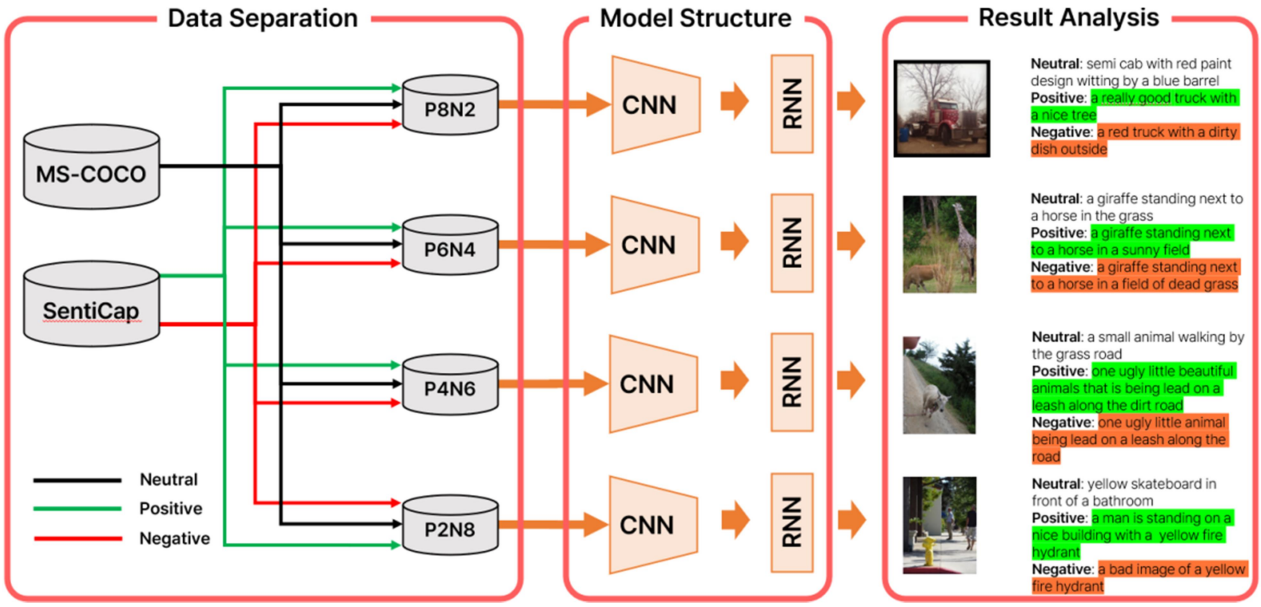


Figure 5. Overall structure of study 2

Figure 5는 두 번째 연구의 전체적인 구조 및 생성 결과에 대한 그림이다. RNN 모델의 기본 형태인 인코더, 디코더 부분에서 인코더를 InceptionV3 모델로 설정하고 디코더는 Bahdanau 어텐션 메커니즘을 사용하여 모델을 설계하였다. Bahdanau 어텐션 메커니즘은 기계 번역 분야에서 문제점으로 지적됐던 번역하고자 하는 문장이 길어질수록 문장 전방의 정보가 사라지는 문제인 RNN의 기억 상실(long term dependency)을 해결한 모델이다. CNN을 통해 추출된 은닉 상태 벡터(hidden state vector)의 차원이 비교적 커 일반적인 RNN 모델을 사용하여 학습을 진행하면 그래디언트 소실(gradient vanishing) 문제가 발생할 수 있으므로 어텐션을 활용하여 학습을 진행하였다.

4.3 Results

결과 분석은 학습 데이터 전처리 과정에서와 같이 총 4가지 그룹으로 나누어 분석을 진행한다. 하나의 그룹 안에 긍정, 중립, 부정의 세 가지 이미지가 존재한다. 각각의 분류들에 대하여 모델이 생성한 긍정 캡션, 중립 캡션, 부정 캡션을 데이터의 캡션과 비교하여 살펴본다. 생성된 캡션에 대해 파이썬의 NLTK Sentiment Intensity Analyzer (SIA) 라이브러리를 이용하여 정량적인 감성 분석을 통해 각 모델이 감정적으로 어느 정도의 편향을 가지는지 분석한다. 결과적으로 데이터의 편향 비율 변화에 따라 모델이 생성하는 캡션의 변화를 살펴본다.

Table 3. Generated captions and SIA results of each images

Figure	Data type	Results
Figure 4. (a)	Generated positive caption	A strong guy with white jacket against the white snow on his snow covered slope
	SIA	[0.191, 0.809, 0.0]
	Generated neutral caption	A man who is performing a trick on a snowboard
	SIA	[0.0, 0.854, 0.146]
	Generated negative caption	A man is riding a snowboard on top of a cold snow covered slope
	SIA	[0.141, 0.859, 0.0]
Figure 4. (b)	Generated positive caption	A really good truck with a nice tree
	SIA	[0.545, 0.455, 0.0]
	Generated neutral caption	Semi cab with red paint design witting by a blue barrel
	SIA	[0.0, 1.0, 0.0]
	Generated negative caption	A red truck with a dirty dish outside
	SIA	[0.0, 0.535, 0.465]
Figure 4. (c)	Generated positive caption	Beautiful street signs are topping off a vandalized stop sign
	SIA	[0.277, 0.567, 0.156]
	Generated neutral caption	A pole has a stop sign and several street signs
	SIA	[0.0, 0.784, 0.216]
	Generated negative caption	The stop sign is always a bad sign and a stupid sign, people should know to stop at a four way intersection without being told
	SIA	[0.0, 0.504, 0.496]

Figure 4와 Table 3을 통해 P8N2 모델이 생성하는 캡션과 각 캡션에 대한 SIA 분석 결과를 살펴볼 수 있다. 각 사진을 바탕으로 모델이 긍정적인 감정의 캡션, 중립적인 감정의 캡션, 부정적인 감정의 캡션을 생성하도록 하였다. 이처럼 생성하도록 한 이유는 전체 테스트 이미지 데이터를 바탕으로 생성한 캡션들의 SIA 분석 결과를 살펴보기 위함에 있다.

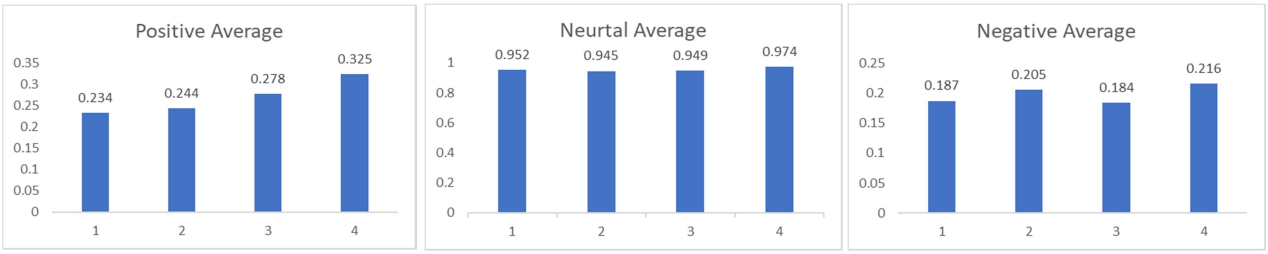
Table 3와 같이 4개의 모델(P8N2, P6N4, P4N6, P2N8)에 대하여 생성되는 캡션을 살펴보고 각 캡션들에 대한 SIA를 수행하였다. SIA 값은 문장의 형용사를 바탕으로 해당 형용사에 긍정, 중립, 부정(Pos, Neu, Neg)의 3가지에 대해 가중치를 미리 설정한 것을 바탕으로 분석된 값이 출력된다. Pos는 문장의 긍정적인 정도, Neu는 중립적인 정도, Neg는 부정적인 정도를 나타내며 모든 값의 합은 1이다. Figure 4 (b)를 바탕으로 생성한 긍정 캡션에 대한 SIA 결과를 살펴보면 [0.545, 0.455, 0.0]로 분석되었다. 이는 상대적으로 부정의 감정이 전혀 담기지 않았고 긍정적인 캡션이라는 것을 의미하게 된다. 또한 부정 캡션에 대한 SIA 결과를 살펴보면 [0.0, 0.535, 0.456]으로, 부정적인 감정이 우세인 문장으로 생성된 캡션임을 알 수 있다.

Table 4는 서로 다른 편향성을 지닌 모델들이 생성한 캡션에 대한 SIA 결과이다. 해당 값들은 Table 3의 SIA 결과값들에 대한 Pos, Neu, Neg의 평균으로, 각 모델의 캡션 생성 성향을 살펴보기 위해 도출하였다. Table 4의 값들을 바탕으로 각 모델의 학습 데이터 편향 정도에 따른 결과의 변화를 보기 위해 그래프들로 구성하였다(Figure 6).

Table 4. NLTK SIA values of each model

Model	Sentiment type	Average values
Positive 80%, Negative 20%	Pos	0.234
	Neu	0.952
	Neg	0.187
Positive 60%, Negative 40%	Pos	0.244
	Neu	0.945
	Neg	0.205
Positive 40%, Negative 60%	Pos	0.278
	Neu	0.949
	Neg	0.184
Positive 20%, Negative 80%	Pos	0.325
	Neu	0.974
	Neg	0.216

Figure 6의 가로축은 각 모델을 나타낸다. 가로축 제일 좌측의 모델이 4개의 모델 중 학습 데이터의 긍정 비율이 가장 높고 부정 비율이 가장 낮은 P8N2 모델이고 가로축 제일 우측의 모델이 4개의 모델 중 학습 데이터의 긍정 비율이 가장 낮고 부정 비율이 가장 높은 P2N8 모델이다. Figure 6. (a) 그래프를 통해 학습 데이터의 긍정 비율이 낮고 부정 비율이 높을수록 모델이 생성하는 캡션의 긍정적인 정도가 높아짐을 확인할 수 있다.



(a) (b) (c)

Figure 6. NLTK SIA graphs of each sentiment type

5. Conclusion

인간과 인간이 상호작용하는 과정에서 감정은 중요한 요소이다. OpenAI의 ChatGPT, Google의 Bard와 같은 생성형 인공지능이 발전하면서 인간과 상호작용이 일상이 되어가고 있다. 이러한 시점에서 본 연구의 주요 목적인 감정이 인공지능 모델의 학습에 어떻게 영향을 미치는지 알아보고자 함은 본 연구뿐만 아니라 앞으로의 연구들에도 중요하게 고려되어야 하는 부분일 것이다.

첫 번째 연구는 CNN visualization을 통해 긍정 데이터만을 이용하여 학습한 CNN 모델과 부정 데이터만을 이용하여 학습한 CNN 모델이 결과를 추론하는 과정에 대한 차이를 살펴보았다. 긍정 감정을 학습한 모델은 사람의 표정을 판단하면서 눈을 주로 판단 요소로

써 활용하고, 부정 감정을 학습한 모델은 입과 그 주변 영역이 결과 예측에 큰 영향을 준다는 것을 확인하였다. 모델이 결과를 판단하는 과정에 차이가 있다는 것은, 모델 내부 구조에 수리적인 차이가 발생하였다는 증거이며 이는 학습 데이터의 감정 요소가 편향되었을 때 인공지능 모델이 결과를 추론하는 과정에 차이가 발생한다는 결론으로 수렴할 수 있다.

두 번째 연구는 첫 번째 연구의 한계점인 단일 감정 데이터를 활용한 학습을 개선하고 데이터의 편향 비율에 따른 결과의 차이를 살펴보기 위해 진행되었다. SIA를 이용하여 서로 다른 비율로 구성한 감정 데이터를 이미지 캡셔닝 모델이 학습하고 이에 따른 결과의 경향성을 살펴보았다. 이미지 캡셔닝 모델 학습에 사용하는 데이터에 긍정과 부정의 감정 요소에 대한 편향 비율의 변화가 발생하였을 때, 긍정에 대한 데이터가 부정에 대한 데이터에 비해 비율적으로 감소하면 모델은 더욱 긍정적인 캡션을 생성하는 것을 확인할 수 있었다.

본 연구는 감정에 대한 데이터의 편향성이 CNN과 이미지 캡셔닝 모델의 추론 과정에 미치는 영향과 모델이 생성하는 결과가 데이터의 편향 정도로 인하여 어떤 변화를 나타내는지 두 가지 연구를 통하여 살펴보았다. 앞서 언급하였듯이 감정이라는 요소는 인간과 상호작용하는 인공지능을 설계할 때 중요하게 고려되어야 하는 부분이다. 이에 착안하여 진행한 본 연구는 감정 데이터가 모델의 구조적 특성에 어떤 변화를 초래하는지 시각적으로 확인함과 동시에 정량적인 분석을 통하여 살펴보았다. 결과적으로 감정을 기준으로 편향된 데이터를 학습한 모델은 긍정과 부정에 따른 편향에 따라 유의미한 차이를 보여주었으며 이는 감정이 인공지능 모델에게 영향을 줄 수 있다는 점을 시사한다.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2022-0-00078, Explainable Logical Reasoning for Medical Knowledge Generation).

References

Al-Moslmi, T., Omar, N., Abdullah, S. and Albared, M., Approaches to cross-domain sentiment analysis: A systematic literature review. *IEEE Access*, 5, 16173-16192, 2017.

Bahdanau, D., Cho, K. and Bengio, Y., Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv: 1409.0473*, 2014.

Barrett, L.F., The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1), 1-23, 2017.

Cagrandi, M., Cornia, M., Stefanini, M., Baraldi, L. and Cucchiara, R., Learning to select: A fully attentive approach for novel object captioning. *Proceedings of the 2021 International Conference on Multimedia Retrieval* (pp. 437-441), 2021.

Chakriswaran, P., Vincent, D.R., Srinivasan, K., Sharma, V., Chang, C.Y. and Reina, D.G., Emotion AI-driven sentiment analysis: A survey, future research directions, and open issues. *Applied Sciences*, 9(24), 5462, 2019.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., Imagenet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248-255), 2009.

Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J. and Zisserman, A., The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111, 98-136, 2015.

Fontaine, J.R., Scherer, K.R., Roesch, E.B. and Ellsworth, P.C., The world of emotions is not two-dimensional. *Psychological Science*, 18(12), 1050-1057, 2007.

Hortensius, R., Hekele, F. and Cross, E.S., The perception of emotion in artificial agents. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4), 852-864, 2018.

Jiang, P.T., Zhang, C.B., Hou, Q., Cheng, M.M. and Wei, Y., Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 30, 5875-5888, 2021.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P., Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324, 1998.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., Microsoft coco: Common objects in context. *Computer Vision-ECCV 2014: 13th European Conference*, Part V 13 (pp. 740-755), 2014.

Mathews, A., Xie, L. and He, X., Senticap: Generating image descriptions with sentiments. *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 30, No. 1), 2016.

Mehrabian, A. and Russell, J.A., An approach to environmental psychology. *the MIT Press*, 1974.

Mollahosseini, A., Hasani, B. and Mahoor, M.H., Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1), 18-31, 2017.

Muthukrishnan, N., Maleki, F., Ovens, K., Reinhold, C., Forghani, B. and Forghani, R., Brief history of artificial intelligence. *Neuroimaging Clinics*, 30(4), 393-399, 2020.

Peter, M. and Ho, M.T., Why we need to be weary of emotional AI. *AI & Society*, 1-3, 2022.

Russell, J.A., Culture and the categorization of emotions. *Psychological Bulletin*, 110(3), 426, 1991.

Santos, M.S., Soares, J.P., Abreu, P.H., Araujo, H. and Santos, J., Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches [Research Frontier]. *IEEE Computational Intelligence Magazine*, 13(4), 59-76, 2018.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision* (pp. 618-626), 2017.

Sharma, H., Agrahari, M., Singh, S.K., Firoj, M. and Mishra, R.K., Image captioning: a comprehensive survey. *International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)* (pp. 325-328), 2020.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z., Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2818-2826), 2016.

Tan, M. and Le, Q., Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning* (pp. 6105-6114), 2019.

Vallverdu, J., Talanov, M., Distefano, S., Mazzara, M., Manca, M. and Tchitchigin, A., NEUCOGAR: A neuromodulating cognitive architecture for biomimetic emotional AI. *International Journal of Artificial Intelligence*, 14(1), 27-40, 2016.

Vinyals, O., Toshev, A., Bengio, S. and Erhan, D., Show and tell: A neural image caption generator. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3156-3164), 2015.

You, Q., Jin, H. and Luo, J., Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions. *arXiv preprint arXiv:1801.10121*, 2018.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A., Learning deep features for discriminative localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2921-2929), 2016.

Author listings

Inheon Choi: ih9511@korea.ac.kr
Highest degree: B.S., School of Electrical Engineering, Hanyang University ERICA
Position title: Graduate Student, Department of Industrial and Management Engineering, Korea University
Areas of interest: Explainable Artificial Intelligence, Deep Learning, Human-AI Interaction, Ethical AI

Sangwon Lee: upcircle@korea.ac.kr
Highest degree: Ph.D., Department of Industrial and Manufacturing Engineering, The Pennsylvania State University
Position title: Professor, Department of Industrial and Management Engineering, Korea University
Areas of interest: User Experience, Human-Computer Interaction, Explainable Artificial Intelligence, Affective Computing, User Modeling