Journal of Korea Multimedia Society Vol. 24, No. 7, July 2021(pp. 880-889)
https://doi.org/10.9717/kmms.2021.24.7.880

균형 잡힌 데이터 증강 기반 영상 감정 분류에 관한 연구

정치윤†, 김무섭††

A Study on Visual Emotion Classification using Balanced Data Augmentation

Chi Yoon Jeong†, Mooseop Kim††

ABSTRACT

In everyday life, recognizing people's emotions from their frames is essential and is a popular research domain in the area of computer vision. Visual emotion has a severe class imbalance in which most of the data are distributed in specific categories. The existing methods do not consider class imbalance and used accuracy as the performance metric, which is not suitable for evaluating the performance of the imbalanced dataset. Therefore, we proposed a method for recognizing visual emotion using balanced data augmentation to address the class imbalance. The proposed method generates a balanced dataset by adopting the random over-sampling and image transformation methods. Also, the proposed method uses the Focal loss as a loss function, which can mitigate the class imbalance by down weighting the well-classified samples. EfficientNet, which is the state-of-the-art method for image classification is used to recognize visual emotion. We compare the performance of the proposed method with that of conventional methods by using a public dataset. The experimental results show that the proposed method increases the F1 score by 40% compared with the method without data augmentation, mitigating class imbalance without loss of classification accuracy.

Key words: Visual Emotion, Emotion Recognition, Data Augmentation, Class Imbalance, Deep Learning

1. 서 론

컴퓨터 비전 분야에서 딥러닝 알고리즘이 발전하면서 객체 탐지, 얼굴 인식 등 전통적인 태스크에서는 사람의 인지 능력을 뛰어넘는 결과를 보여주고 있다[1,2,3]. 따라서, 컴퓨터 비전 분야에서는 새로운 태스크를 찾기 위한 다양한 시도가 계속되고 있다[4]. 최근 소셜네트워킹 서비스, 미디어 공유 서비스 등 멀티미디어 중심의 소셜네트워크 플랫폼이 증가하면서, 사용자들이 텍스트 대신 영상과 동영상을 통하여 자신의 감정을 표현하는 사례가 늘어나고 있다. 따라서, 영상을 통하여 표출되는 감정을 인식하는 영상 감정 분류에 관한 관심이 증가하고 있다. 영상 감정 분류는 감정이 가지는 모호성으로 인하여 많은 연구 이슈들이 존재하고, 인간-컴퓨터 상호 작용, 영상 감시, 로보틱스, 게임, 엔터테인먼트 등 다양한 분야에 적용 가능하여 학계 및 산업계에서 많은 관심을 받고 있다[5,6].

※ Corresponding Author : Mooseop Kim, Address: (34129) 218 Gajeong-ro, Yuseong-gu, Daejeon, Korea, TEL : +82-42-860-1340, FAX : +82-42-860-5545, E-mail : gomskim@ etri.re.kr
Receipt date : May 12, 2021, Revision date : Jul. 9, 2021
Approval date : Jul. 13, 2021
† Human Enhancement & Assistive Technology Research Section, Artificial Intelligence Research Lab., ETRI (E-mail : iamready@etri.re.kr)
†† Human Enhancement & Assistive Technology Research Section, Artificial Intelligence Research Lab., ETRI
※ This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. (21ZS1200, Fundamental Technology Research for Human-Centric Autonomous Intelligent Systems)

영상 감정 분류의 초기 연구들[7,8]은 전문가가 정의한 특징(handcrafted features)과 기계학습 방법을 사용하였으며, 최근에는 딥러닝 알고리즘을 사용하여 영상 감정을 분류하는 연구들이 진행되고 있다[4, 9,10,11,12]. 딥러닝 기반 연구들은 영상 분류 분야에서 많이 사용되는 딥러닝 모델을 사용하여 성능을 개선하거나[9], 계층적으로 분류할 수 있는 감정의 특성을 활용하여 커리큘럼 학습(curriculum learning)을 적용함으로써 분류 성능을 개선하고자 하였다[4]. 또한, 영상 감정은 영상 속에 존재하는 사람의 표정, 자세, 행동 등에 영향을 받기 때문에 전체 영상 정보와 사람의 특성정보를 통합하여 분석하는 방법[10,11] 및 영상과 텍스트 정보를 통합하여 분석하는 방법[12] 등에 관한 연구가 진행되었다.

음성 및 생체 신호를 기반으로 사람의 감정을 분석하는 기존 연구들을 살펴보면, 사람의 감정은 다양한 감정 범주에 균일하게 분포하지 않고 불균형이 심한 특성이 있다[13]. 이와 유사하게 영상 감정 정보 또한 소수의 특정 감정에 데이터가 집중되어 분포하는 특성이 있으며, 최근 공개된 SE30K8 데이터셋[12]의 감정 범주별 분포에서도 이러한 특성을 확인할 수 있다. Fig. 1은 SE30K8 데이터셋의 8가지 감정 범주에 대한 데이터 수를 나타내며, 특정 감정에 데이터가 집중된 것을 확인할 수 있다. 범주 불균형을 측정하는 지표인 불균형 비율(imbalance ratio)[14]은 가장 적은 데이터를 가지는 범주와 가장 많은 데이터를 가지는 범주의 데이터 비율로 계산되며, SE30K8 데이터셋의 경우 불균형 비율이 54 정도로 불균형이 심하다는 것을 확인할 수 있다. 하지만, 기존 연구들은 이러한 범주 불균형에 대한 문제를 다루지 않고 있으며, 영상 감정 분류 성능을 평가하는 지표로 정확도(accuracy)를 사용하여 분류 성능이 왜곡되는 문제점이 있었다.

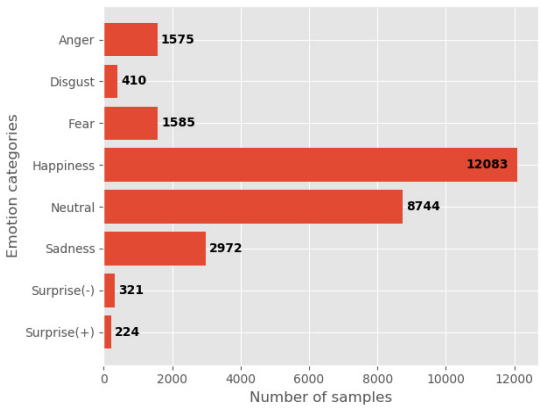

Fig. 1. Details of the SE30K8 dataset.

따라서, 본 논문에서는 균형 잡힌 데이터 증강을 이용하여 영상 감정 분류 성능을 향상할 수 있는 새로운 방법을 제안하였다. 먼저, 본 논문에서는 영상 감정 데이터셋의 범주 불균형을 해소하기 위하여 데이터 관점에서 균형 잡힌 학습 데이터를 생성하는 방법을 제안하였다. 또한, 알고리즘 관점에서 범주 불균형을 해소하기 위해서 제안된 손실함수(loss function)인 Focal loss[15]를 분류 네트워크 모델을 학습하는 데 사용하였다. 영상 감정을 분류하는 네트워크 모델로는 영상 인식 분야에서 뛰어난 성능을 보여주고 EfficientNet[16]을 사용하였으며, 분류 성능의 정확한 평가를 위하여 기존 연구에서 사용한 정확도에 F1 스코어(F1 score)를 추가하여 측정지표로 사용하였다. 본 논문에서 제안한 방법을 공개 데이터셋을 사용하여 분류 성능을 측정한 결과, 제안 방법이 기존 방법보다 향상된 성능을 가지는 것을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서 영상 감정 인식 방법 및 범주 불균형 해소를 위한 연구 동향에 관하여 살펴보고, 3장에서는 본 논문에서 제안하는 균형 잡힌 데이터 증강 기반 영상 감정 분류방법을 소개한다. 4장에서는 공개 데이터셋을 사용하여 제안 방법을 실험한 결과를 객관적인 지표로 비교하여 성능을 평가하고, 5장에서는 결론 및 향후 연구 방향을 제시한다.

2. 관련 연구

2.1 영상 감정 분류 기술

영상 감정을 분류하는 초기 방법들은 전문가가 정의한 특징을 기계학습 알고리즘을 사용하여 분류하는 방법[7,8]들이 주로 연구되었으며, 최근에는 딥러닝을 사용하여 영상 감정을 분류하는 연구들이 진행되고 있다[4,9,10,11,12]. 딥러닝 기반 연구들은 딥러닝 모델의 구조 및 학습 방법을 개선하여 성능을 향상하는 방법 및 영상 정보와 다른 부가 정보를 통합하여 성능을 향상하는 방법에 초점을 맞추고 있다.

기존 연구[9]에서는 이미지넷 데이터를 사용하여

사전 학습된 AlexNet[17]에 전이학습(transfer learning)을 적용하여 영상 감정을 분류하는 방법을 제안하였다. 제안된 방법의 성능 검증을 위하여 8가지 감정에 대하여 23,000장 규모의 데이터셋을 크라우드소싱 플랫폼을 사용하여 구축하였다. 구축된 데이터셋에 전이학습을 통하여 생성된 분류 모델을 적용한 결과 약 58%의 분류 정확도를 보여주었다.

기존 영상 감정 데이터셋들은 학습 데이터의 양이 제한적이어서 데이터의 편향성이 발생하기 때문에, 대용량 데이터셋을 활용하여 일반화 성능을 향상하기 위한 연구[4]가 진행되었다. 제안 방법에서는 쉽게 분류할 수 있는 단순한 문제로 시작하여 점차 복잡한 문제를 학습시킴으로써 딥러닝 모델의 분류 성능을 향상하는 커리큘럼 학습을 적용하였다. 긍정과 부정의 2개 범주로 구분된 감정 데이터로부터 학습을 시작하여 점점 더 세분된 감정을 분류하도록 학습하였으며, ResNet-50 모델[18]을 사용하여 실험한 결과 26개의 감정 범주를 한 번에 학습하는 경우보다 정확도가 약 3% 정도의 향상되는 결과를 보여주었다.

영상 감정은 영상에 존재하는 사람과 주변 상황 정보에 많은 영향을 받기 때문에 두 개의 정보 추출 네트워크를 사용하여 사람에 대한 정보와 영상 전체에 대한 정보를 분석하고, 이를 통합하여 영상 감정을 분류하는 방법이 제안되었다[10]. 제안 방법에서는 영상에서 검출된 사람의 영역 정보를 활용하였으며, 전체 영상 정보만을 사용하는 경우보다 사람의 영역 정보를 활용하면 평균 정밀도(average precision)가 약 4% 정도 향상되는 결과를 보여주었다. 또한, 컬러 영상 정보로부터 사람의 얼굴, 포즈, 배경 정보를 분석하고, 영상의 깊이 정보로부터 객체 간의 관계 정보를 분석하여 영상 감정을 분류하는 방법도 제안되었다[11].

최근에는 영상 정보와 텍스트 정보를 함께 사용하여 웹 데이터의 감정을 분석하기 위한 연구가 진행되었다[12]. 이 연구에서는 백만 장 규모의 영상 및 텍스트 정보로 구성된 StockEmotion 데이터셋을 구축했으며, 한 장의 영상에 평균 48개의 키워드, 7개의 감정 관련 키워드로 구성되어 있다. 전체 데이터셋 중 사람의 검증을 통하여 8개의 감정 범주로 구분한 SE30K8 데이터셋을 구축하고, 영상 정보와 텍스트 정보를 모두 활용한 딥러닝 모델을 사용하는 경우 영상 정보만을 사용하는 경우보다 정확도가 17% 정도 향상되는 것을 보여주었다.

2.2 범주 불균형 완화 기술

영상 감정 데이터의 경우 소수의 특정 감정 범주에 데이터가 집중되어 범주 불균형이 발생하기 때문에 이를 완화하는 방법이 필요하다. 딥러닝에서 범주 불균형을 해결하는 방법은 데이터 관점의 접근 방법과 알고리즘 관점의 접근 방법으로 구분된다[14].

데이터 관점의 접근 방법은 랜덤 언더 샘플링(random undersampling)을 통하여 다수의 데이터를 갖는 범주의 데이터를 선택적으로 추출하여 데이터 수를 줄임으로써 균형을 맞추는 방법[19]과 랜덤 오버 샘플링(random oversampling, ROS)을 통하여 소수의 데이터를 복사하여 해당 범주의 데이터 수를 증가시키는 방법[20] 및 실시간으로 범주별 영상의 분류 성능을 분석하여 샘플 사이즈를 조정하는 방법[21] 등이 있다.

알고리즘 관점의 접근 방법은 주로 딥러닝 학습에 사용되는 손실함수를 수정하여 소수 범주의 데이터에 가중치를 부여하는 방법이 많이 활용된다. 대표적으로 사용되는 손실함수로는 중심 손실(focal loss)[15]과 평균 오류 오차 손실(mean false error loss)[22] 등이 있다.

3. 균형 잡힌 데이터 증강 기반 영상 감정 분류 방법

본 논문에서 제안하는 균형 잡힌 데이터 증강 기반 영상 감정 분류방법은 Fig. 2와 같이 균형 잡힌 학습 데이터셋을 생성하는 단계와 생성된 데이터셋으로부터 딥러닝 모델과 Focal loss를 사용하여 분류

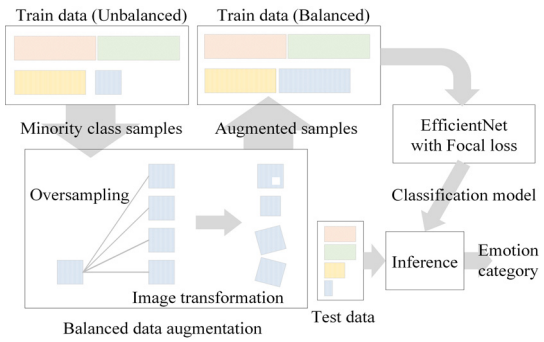

Fig. 2. Flowchart of the proposed method.

모델을 생성하는 단계 및 생성된 분류 모델을 사용하여 영상 감정을 분류하는 단계로 구성된다.

균형 잡힌 학습 데이터셋을 생성하는 단계에서는 학습 데이터의 범주 별 샘플 수를 분석하여 데이터의 양이 적은 범주에 대해서 ROS을 수행한다. ROS는 기존 데이터 중 랜덤하게 샘플을 선택하여 데이터셋에 추가함으로써 데이터양을 증가시키는 방법이다[14]. 오버 샘플링을 수행하게 되면 중복된 데이터가 다수 존재하여 과적합(over-fitting)이 발생할 수 있다. 따라서 제안 방법에서는 오버 샘플링된 데이터에 대해서 데이터 증강 방법에서 사용되는 영상 변형을 적용함으로써 과적합을 방지하였다.

데이터 증강 방법에는 다양한 영상 변형 방법들이 적용될 수 있으며, 최근 연구 결과에 따르면 적용 도메인에 따라서 최적의 영상 변형 방법이 달라진다고 알려져 있다[23,24]. 따라서 본 논문에서는 먼저 영상 감정 데이터셋에 다양한 영상 변형을 적용하여 성능 향상 정도를 분석한 후, 성능 향상이 큰 5개의 영상 변형 방법을 ROS 결과로 생성된 데이터에 적용하여 균형 잡힌 데이터셋을 생성하였다. 제안 방법은 소수의 데이터를 가지는 범주에 대해서만 오버 샘플링을 수행하고, 영상 감정 분야에서 성능 향상 정도가 큰 영상 변형을 적용함으로써 ROS 방법이 가지는 과적합 문제를 해결하였다. 또한, 모든 범주의 데이터 수를 동일한 비율로 증가시키는 단순 데이터 증강 방법과 달리 범주별 균형 잡힌 데이터셋을 생성할 수 있는 장점이 있으며, 이로 인하여 영상 감정 분류 성능이 향상될 것으로 기대된다.

기존 영상 감정 분류 연구들[4,9,10,11,12]은 AlexNet[17], ResNet[18]과 같이 전문가가 설계한 네트워크 모델을 주로 사용하였다. 최근 태스크에 적합한 최적의 네트워크 모델을 자동으로 생성하는 방법에 관한 연구가 진행되면서, 기계가 생성한 네트워크 모델이 기존 모델들의 성능을 뛰어넘고 있다[16]. 네트워크 모델의 성능은 네트워크의 깊이 및 넓이, 입력 영상의 해상도가 균형을 맞춰 변화할 때 성능이 향상된다는 것을 확인하고, 이들의 상관관계를 고려하여 최적의 네트워크 구조를 생성한 EfficientNet[16]이 제안되었다. EfficientNet은 입력 영상의 해상도 및 모델 크기에 따라서 EfficientNet-B0 ~ B7까지 8개의 모델이 존재하며, EfficientNet-B0의 경우 이미지넷 분류 정확도에서 ResNet-50, DenseNet-169 모델보다 더 작은 모델 크기로 더 높은 성능을 보여주었다[16]. 따라서 본 논문에서는 영상 감정 분류 네트워크로 EfficientNet을 사용하여 모델별 영상 감정 분류 성능을 분석하였다.

네트워크 모델을 학습할 때 손실함수로는 교차 엔트로피(cross entropy)가 일반적으로 사용된다. 하지만, 교차 엔트로피의 경우 범주별 데이터의 분포와 관계없이 모든 샘플들이 손실 값에 미치는 영향이 동일하여, 범주 불균형이 심한 경우 다수 데이터를 갖는 범주의 샘플만을 잘 구별하기 위한 방향으로 학습되는 문제점이 있다. 따라서 제안 방법에서는 데이터가 적게 분포하는 범주의 샘플에 가중치를 주기 위해서 제안된 Focal loss[15]를 손실함수로 사용하였으며, Focal loss는 다음 식 (1)과 같이 정의된다.

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \quad (1)$$

위의 식에서 p_t는 해당 범주를 정확하게 분류할 확률이며, α_t는 가중치, γ는 집중 정도를 설정하는 변수이다. 데이터셋에서 다수 샘플을 가지는 범주의 경우 해당 범주를 정확하게 분류할 확률이 높아지기 때문에 손실에 미치는 영향이 적어진다. 반대로, p_t가 낮은 경우에는 손실 값에 미치는 영향을 많이 증가시킴으로써 소수를 차지하는 범주의 샘플들을 잘 분류할 수 있도록 학습하게 된다. 제안 방법에서 가중치 α_t는 0.25, 집중 정도 γ는 2.0을 사용하였다.

본 논문에서 제안하는 균형 잡힌 데이터 증강 기반 영상 감정 분류 모델의 학습 방법은 Fig. 3의 알고리즘과 같다. 먼저 학습 데이터의 범주별 샘플 수를 분석하여, 최소 샘플 수(N_{min})보다 적은 경우에는 6-9 번째 라인과 같이 ROS를 수행한다. ROS를 통하여 선택된 샘플 데이터에 대해서 11-14번째 라인과 같이 다양한 영상 변형 방법을 적용하며, 이때 영상 변형 방법의 집합(T)은 사전에 성능 분석을 통하여 성능 향상이 큰 5개의 영상 변형 방법으로 구성된다. 균형 잡힌 데이터셋이 생성되면 이미지넷 데이터로 사전 학습된 EfficientNet과 Focal loss를 사용하여 전이학습을 수행함으로써 영상 감정 분류 모델을 생성한다. 본 논문에서 제안한 방법은 범주 불균형을 해소하기 위한 데이터 관점의 접근 방법으로 균형 잡힌 학습 데이터셋을 생성하고, 알고리즘 관점의 접근 방법으로 Focal loss를 조합해서 사용함으로써 범주 불균형이 심한 영상 감정 데이터의 분류 성능을

Algorithm 1: Model training with balanced data augmentation

Input: Training images X_{train}, class list C, the minimum number of samples N_{min}, image transformation list T,
Output: Visual emotion classification model M

```
/* Initialization                                                    */
A ← ∅
/* Generate the balanced training images                             */
for each class c_j ∈ C do
    N_j ⟵ CalculateSampleSize(X_train, c_j)
    S_j ⟵ GetClassSample(X_train, c_j)
    /* Random over-sampling                                          */
    A_m ← ∅
    while N_j < N_min do
        s ⟵ GetRandomSample(S_j)
        A_m ⟵ AddSampleImage(A_m, s)
        N_j ⟵ N_j + 1
    end
    /* Image transformation                                          */
    for each sample a_i ∈ A_m do
        t ⟵ SelectImageTransformation(T)
        a'_i ⟵ ApplyImageTransformation(a_i, t)
        A ⟵ AddSampleImage(A, a'_i)
    end
end
/* Model training                                                    */
M ⟵ TrainModel(X_train, A)
```

Fig. 3. Proposed algorithm for model training with balanced data augmentation.

향상할 것으로 기대된다.

4. 실험 결과

균형 잡힌 데이터 증강 기반 영상 감정 분류 성능 분석을 위하여 8개의 감정 범주로 수집된 SE30K8 데이터셋[12]을 사용하였으며, 감정 범주별 데이터의 구성은 Fig. 1과 같이 약 28,000장의 이미지로 구성되어 있다. Fig. 4는 SE30K8 데이터셋의 감정 범주별 샘플 영상을 나타낸다.

본 연구에서는 EfficientNet 기반의 제안 방법과 기존 연구에서 가장 많이 사용된 ResNet 모델[18]과 VGG 모델[25]을 사용하여 네트워크 모델에 따른 영상 감정 분류 성능 차이를 분석하였다. 또한, 네트워크 모델 학습 시 손실함수에 따른 분류 성능을 분석하기 위하여 교차 엔트로피와 Focal loss를 손실함수로 사용하여 성능을 비교하였으며, 분류 성능 측정을

Fig. 4. Visual examples of the 8 emotion categories in the SE30K8 Dataset.

위한 평가척도로 정확도와 F1 스코어를 모두 사용하였다. 영상 감정 분류 성능 측정은 5겹 교차 검증(5-fold cross validation) 결과의 평균값을 사용하였다. 영상 감정 분류방법들은 Keras 라이브러리를 사용하여 구현되었으며, 성능 측정은 Intel Xeon W-2295 프로세스와 Quadro RTX 8000 그래픽 카드가 탑재된 윈도우 10 64bit 환경에서 수행되었다.

최근 연구[23,24]에서 적용 도메인에 따라서 최적의 영상 변형 방법이 달라진다고 알려져 있어, 본 논문에서는 먼저 단순 데이터 증강 방법을 적용했을 때 영상 변형 방법별 성능 향상 정도를 분석하였다. 데이터 증강은 학습 데이터에 대해서 한 번씩 영상 변형을 적용하였으며, 그 결과 학습 데이터의 양은 2배로 증가하였다. 데이터 증강 방법에 사용된 상세한 방법 및 설정값은 Table 1과 같다.

또한, 데이터 증강 방법 적용 시 네트워크 분류 모델과 손실함수에 따른 영상 감정 분류 성능 변화를 실험하였다. 네트워크 모델 학습을 위한 최적화 방법으로는 ADAM(adaptive moment estimation)을 사용하였으며, 학습 비율은 0.0001, 배치 사이즈는 128, 반복(epoch) 횟수는 30으로 설정하였다.

데이터 증강 방법과 딥러닝 네트워크 모델 및 손실함수에 따른 영상 감정 분류 성능은 Table 2 및 Table 3과 같다. Table 2는 손실함수로 교차 엔트로피를 사용했을 때 데이터 증강 방법과 네트워크 모델에 따른 영상 감정 분류 성능을 나타내며, Table 3은

Table 1. List of all image transformations.

Operation Name	Description	Range of magnitudes
Blur	Apply the Gaussian blur using magnitude-sized kernel	[3, 7]
Random crop	Randomly crop the image to target height and width	[200, 200]
Random flip	Randomly flip the image horizontally and vertically	-
Rotation	Randomly rotate the image the rate magnitude of 2π	[-0.125, 0.125]
Translation	Randomly translate the image the rate magnitude	[-0.3, 0.3]
Cutout[26]	Set a random square patch of side-length magnitude pixels to gray	[0, 70]
Random contrast	Control the contrast of the image by rate magnitude	[-0.3, 0.3]
Random height	Randomly vary the height of image by rate magnitude	[-0.3, 0.3]
Random width	Randomly vary the width of image by rate magnitude	[-0.3, 0.3]

Table 2. Results of visual emotion classification according to network models using cross entropy loss and various data augmentation methods.

	Cross Entropy Loss					
	VGG-16		ResNet-50		EfficientNet-B0	
	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
No augmentation	54.36%	0.189	56.82%	0.257	57.86%	0.261
Blur	54.69%	0.198	56.92%	0.266	58.15%	0.274
Random crop	54.94%	0.199	57.33%	0.271	58.22%	0.274
Random Flip	54.69%	0.189	57.28%	0.262	58.21%	0.274
Rotation	54.94%	0.200	57.21%	0.263	58.07%	0.268
Translation	55.06%	0.201	57.15%	0.271	58.14%	0.274
Cutout	54.84%	0.200	57.14%	0.262	58.11%	0.271
Random contrast	54.97%	0.191	57.31%	0.273	58.44%	0.275
Random height	54.75%	0.197	57.18%	0.269	58.37%	0.276
Random width	55.07%	0.199	57.16%	0.269	58.17%	0.276
Random zoom	54.85%	0.198	57.23%	0.271	58.41%	0.275

Table 3. Results of visual emotion classification according to network models using Focal loss and various data augmentation methods.

	Focal loss					
	VGG-16		ResNet-50		EfficientNet-B0	
	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
No augmentation	54.35%	0.185	56.76%	0.248	57.68%	0.249
Blur	55.19%	0.206	57.30%	0.277	58.24%	0.275
Random crop	55.36%	0.208	57.22%	0.280	58.35%	0.287
Random Flip	55.18%	0.198	57.10%	0.263	57.92%	0.272
Rotation	55.27%	0.201	57.19%	0.277	58.25%	0.280
Translation	55.18%	0.205	57.16%	0.280	58.27%	0.279
Cutout	55.18%	0.202	57.01%	0.269	58.35%	0.280
Random contrast	55.20%	0.202	57.44%	0.280	57.94%	0.280
Random height	55.41%	0.202	57.28%	0.283	58.04%	0.272
Random width	55.13%	0.201	57.28%	0.276	58.37%	0.283
Random zoom	55.18%	0.206	57.25%	0.271	58.36%	0.281

손실함수로 Focal loss를 사용했을 때 영상 감정 분류 성능을 나타낸다. Table 2와 Table 3의 결과를 분석해보면 EfficientNet-B0를 사용하는 경우 손실함수 및 데이터 증강 방법에 무관하게 기존 네트워크 모델보다 뛰어난 성능을 보여주었다. 또한, 데이터 증강 방법을 적용하는 경우 모든 네트워크 모델에서 성능이 향상되는 결과를 보여주었다. 손실함수는 Focal loss를 사용하는 경우 교차 엔트로피를 사용하는 경우보다 대부분 높은 성능을 보여주었으며, 특히 F1 스코어 측면에서 Focal loss가 더 좋은 성능을 보여주었다. EfficientNet-B0의 F1 스코어를 기준으로 랜덤 플립과 랜덤 높이를 제외한 모든 변형 방법에서 Focal loss가 교차 엔트로피보다 높은 성능을 보여주었다. 따라서, 본 논문에서는 EfficientNet-B0에서 F1 스코어가 높은 5가지 영상 변형 방법, 랜덤 크롭, 회전, Cutout, 랜덤 폭, 랜덤 줌을 균형 잡힌 데이터 생성을 위한 영상 변형 방법으로 선택하였다.

또한, 본 연구에서는 데이터 증강을 적용하지 않은 경우의 영상 감정 분류 성능을 기준으로, ROS만 적용한 경우와 ROS 적용 후 10가지 영상 변형 방법을 적용한 경우 및 ROS 적용 후 본 논문에서 도출한 최적의 영상 변형 방법들을 적용한 제안 방법의 성능을 각각 비교하여 데이터 증강의 효율성을 검증하도록 실험을 확장하였다. ROS는 범주별 데이터 수가 최소 샘플 수(N_{min})보다 적은 경우에 적용되며, 실험에서는 N_{min}을 1500으로 설정하였다. 영상 변형 방법의 집합(T)은 이전 실험에서 높은 성능을 보인 5가지 영상 변형 방법으로 구성되며, 이전 실험과 동일하게 네트워크 모델 학습을 위한 최적화 방법으로 ADAM을 사용하였다. 학습 비율은 0.0001, 배치 사이즈는 128로 설정하였으며, 학습의 반복 횟수는 300으로 설정한 후, 10번 학습을 반복하는 동안 성능 변화가 없는 경우에 조기 종료하도록 설정하였다.

EfficientNet-B0를 사용한 제안 방법과 다른 방법들의 SE30K8 데이터셋에 대한 영상 감정 분류 성능은 Table 4와 같다. Table 4를 살펴보면 오버 샘플링을 적용하는 경우 범주 불균형이 존재할 때 중요한 평가척도가 되는 F1 스코어가 향상되는 것을 보여주었다. 또한, 오버 샘플링된 데이터에 이전 실험에서

Table 4. Results of visual emotion classification based on SE30K8 dataset.

	EfficientNet-B0 (Focal loss)	
	Accuracy	F1 score
Baseline	57.68%	0.249
Over-sampling	56.81%	0.319
Over-sampling + all image transformations	57.13%	0.319
Proposed method	57.18%	0.325

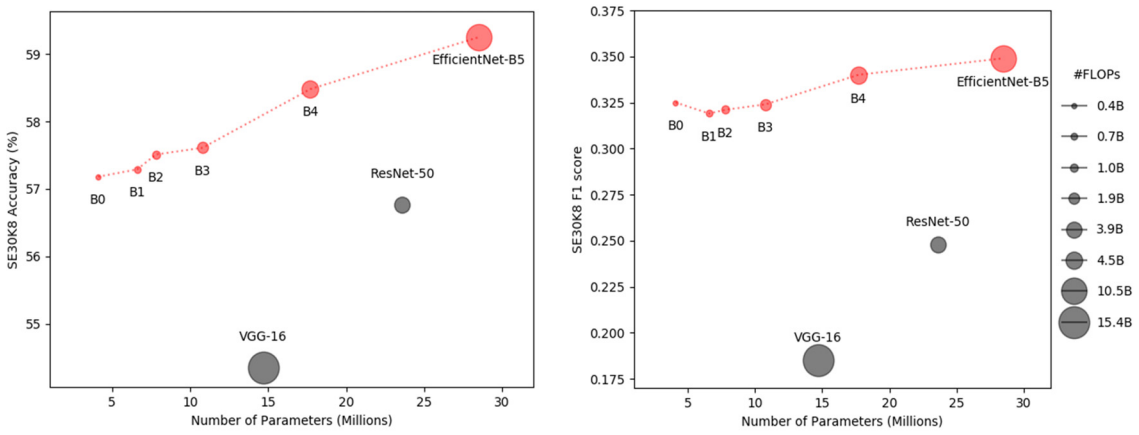

Fig. 5. Performance results on SE30K8 dataset. Red circles indicate the proposed methods that use balanced data augmentation and black circles indicate the existing methods that do not use data augmentation.

도출한 최적의 영상 변형 방법을 적용하여 균형 잡힌 데이터셋을 생성하고, 이를 사용하여 학습하는 경우 F1 스코어 및 정확도가 오버 샘플링만 수행한 경우보다 모두 증가하였다. 제안 방법은 데이터 증강을 적용하지 않은 기준 성능보다 정확도는 0.5% 낮아지지만, 범주 불균형이 존재할 때 중요한 평가 요소인 F1 스코어는 0.325로 기준 성능보다 30% 정도 향상되는 결과를 보여주었다.

Fig. 5는 제안 방법의 분류 모델을 EfficientNet-B0부터 B5로 변경하였을 때 영상 감정 분류 성능과 기존 방법의 성능을 비교한 결과를 나타낸다. Fig. 5에서는 정확도와 F1 스코어에 따른 네트워크 모델의 파라미터 수와 FLOPs(floating point operations)를 비교하였으며[27], EfficientNet을 사용하는 경우 기존 모델보다 적은 연산량으로 더 높은 성능을 보여주는 것을 확인할 수 있다. 또한, EfficientNet-B0에서 B5로 모델의 복잡도가 높아지는 경우 정확도 및 F1 스코어가 점진적으로 증가하는 것을 확인할 수 있다. EfficientNet-B5를 사용하는 경우 정확도는 59.24%, F1 스코어는 0.348로써 기준 성능 대비 F1 스코어가 약 40% 향상되는 결과를 보여준다.

5. 결 론

본 논문에서는 균형 잡힌 데이터 증강을 이용하여 영상 감정 분류 성능을 향상하는 방법을 제안하였다. 제안 방법은 학습 데이터에 오버 샘플링을 수행한 후, 오버 샘플링된 샘플에 대해서 영상 감정 분류에 적합한 영상 변형 방법들을 도출하여 적용함으로써 균형 잡힌 학습 데이터셋을 생성하였다. 또한, 영상 인식 분야에서 높은 성능을 보여주는 최신 네트워크 모델인 EfficientNet과 범주 불균형을 완화 시킬 수 있는 Focal loss를 손실함수로 사용하여 영상 감정 분류 네트워크를 학습하였다. 제안 방법을 SE30K8 공개 데이터셋에 적용하여 실험한 결과 데이터 증강을 적용하지 않은 경우에 비해서 정확도와 F1 스코어에 모두 향상되는 결과를 보여주었다. 특히 제안 방법은 EfficientNet-B0, B5를 사용할 때 범주 불균형이 존재할 때 중요한 평가 지표인 F1 스코어를 데이터 증강을 적용하지 않은 기준 성능보다 정확도의 손실 없이 각각 30%, 40% 향상시키는 결과를 보여주어 범주 불균형을 효과적으로 완화할 수 있음을 확인하였다.

본 논문에서는 영상 데이터 증강을 위하여 영상처리 기반 변형 방법만을 고려하였지만, 최근 적대적 생성 신경망(generative adversarial network) 기반의 영상 증강 방법에 관한 연구들이 진행되고 있다. 따라서 균형 잡힌 데이터셋 생성을 위한 데이터 증강의 한 방법으로 적대적 생성 신경망을 활용하는 방안에 관한 연구가 필요하다. 또한, 최근 사람의 개입 없이 적용 도메인에 맞는 최적의 영상 증강 방법을 탐색하는 연구들이 진행되고 있어, 향후 연구에서는 자동 기계 학습(autoML)을 적용하여 영상 감정 분야에 적합한 최적의 영상 증강 방법을 도출하고 이를

활용한다면 영상 감정 분류 성능을 개선할 수 있을 것으로 기대된다.

REFERENCE

[1] S. Dodge and L. Karam, "A Study and Comparison of Human and Deep Learning Recognition Performance under Visual Distortions," *Proceedings of the International Conference on Computer Communications and Networks*, pp. 1-7, 2017.

[2] H. Bae and O. Kwon, "Untact Face Recognition System Based on Super-Resolution in Low-Resolution Images," *Journal of Korea Multimedia Society*, Vol. 23, No. 3, pp. 412-420, 2020.

[3] J. Kim, S. Jung, and C. Sim, "A Study on Object Detection using Restructured Retina Net," *Journal of Korea Multimedia Society*, Vol. 23, No. 12, pp. 1531-1539, 2020.

[4] R. Panda, J. Zhang, H. Li, J.-Y. Lee, X. Lu, and A. K. Roy-Chowdhury, "Contemplating Visual Emotions: Understanding and Overcoming Dataset Bias," *Proceedings of the European Conference on Computer Vision*, pp. 594-612, 2018.

[5] W. Hua, F. Dai, L. Huang, J. Xiong, and G. Gui, "Hero: Human Emotions Recognition for Realizing Intelligent Internet of Things," *IEEE Access*, Vol. 7, pp. 24321-24332, 2019.

[6] T. Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li, "Spatial-temporal Recurrent Neural Network for Emotion Recognition," *IEEE Transactions on Cybernetics*, Vol. 49, No. 3, pp. 839-847, 2019.

[7] L. Chang, Y. Chen, F. Li, M. Sun, and C. Yang, "Affective Image Classification Using Multi-Scale Emotion Factorization Features," *Proceedings of the International Conference on Virtual Reality and Visualization*, pp. 170-174, 2016.

[8] J. Cao, Y. Li, and Y. Tian, "Emotional Modelling and Classification of a Large-scale Collection of Scene Images in a Cluster Environment," *PLos ONE*, Vol. 13, No. 1, pp. 1-20, 2018.

[9] Q. You, J. Luo, H. Jin, and J. Yang, "Building a Large Scale Dataset for Image Emotion Recognition: the Fine Print and the Benchmark," *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 308-314, 2016.

[10] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context Based Emotion Recognition Using EMOTIC Dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, pp. 2755-2766, 2020.

[11] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emoti Con: Context-Aware Multimodal Emotion Recognition using Frege's Principle," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14234-14243, 2020.

[12] Z. Wei, J. Zhang, Z. Lin, J.-Y. Lee, N. Balasubramanian, M. Hoai, and D. Samaras, "Learning Visual Emotion Representations from Web Data," *Proceedings of the IEEE/CVF Conference on Computer Vision*, pp. 13106-13115, 2020.

[13] K.J. Noh, C.Y. Jeong, J. Lim, S. Chung, G. Kim, J.M. Lim, and H. Jeong, "Multi-Path and Group-Loss-Based Network for Speech Emotion Recognition in Multi-Domain Datasets," *Sensors*, Vol. 21, pp. 1-18, 2021.

[14] J.M. Johnson and T.M. Khoshgoftaar, "Survey on Deep Learning with Class Imbalance," *Journal of Big Data*, Vol. 6, pp. 1-54, 2019.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," *Proceedings of the IEEE/CVF Conference on Computer Vision*, pp. 2980-2988, 2017.

[16] M. Tan and Q.V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *Proceedings of the International Conference on Machine Learning*, pp. 6105-

6114, 2019.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Proceedings of the Conference on Neural Information Processing Systems*, pp. 1097-1105, 2012.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[18] H. Lee, M. Park, and J. Kim, "Plankton Classification on Imbalanced Large Scale Database via Convolutional Neural Networks with Transfer Learning," *Proceedings of the IEEE International Conference Image Processing*, pp. 3713-3717, 2016.

[19] S. Pouyanfar, Y. Tao, A. Mohan, A.S. Kaseb, K. Gauenn, R. Dailey, and et. al., "Dynamic Sampling in Convolutional Neural Networks for Imbalanced Data Classification," *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval*, pp. 112-117, 2018.

[20] P. Hensman and D. Masko, *The Impact of Imbalanced Training Data for Convolutional Neural Networks*, Bachelor's Thesis of KTH Royal Institute of Technology, pp. 1-28, 2015.

[21] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P.J. Kennedy, "Training Deep Neural Networks on Imbalanced Datasets," *Proceedings of the International Joint Conference on Neural Networks*, pp. 4368-4374, 2016.

[22] S. Porcu, A. Floris, and L. Atzori, "Evaluation of Data Augmentation Techniques for Facial Expression Recognition Systems," *Electronics*, Vol. 9, No. 11, pp. 1-12, 2020.

[23] E.D. Cubuk, B. Zoph, D. Mane, V. Vasudevna, and Q.V. Le, "AutoAugment: Learning Augmentation Strategies from Data," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 113-123, 2019.

[24] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *Proceedings of the International Conference on Learning Representations,* pp. 1-14, 2015.

[25] T. DeVries and G.W. Taylor, "Improved Regularization of Convolutional Neural Networks with Cutout," *arXiv Preprint*, arXiv: 1708.04552, pp. 1-8, 2017.

[26] FLOPs calculator for neural network architecture(2020). https://github.com/tokusumi/keras-flops (accessed June 24, 2021).

정 치 윤

2002년 POSTECH 전자전기공학과 학사
2004년 POSTECH 전자컴퓨터공학부 석사
2018년 KAIST 전산학부 박사
2004년~현재 한국전자통신연구원 책임연구원

관심분야 : Computer vision, Pattern recognition, Machine learning, Sensory substitution

김 무 섭

1996년 경북대학교 전자공학과 학사
1998년 경북대학교 전자공학과 석사
1998년~1999년 LG종합기술원 연구원

2008년 충남대학교 컴퓨터공학과 박사
1999년~현재 한국전자통신연구원 책임연구원
관심분야 : Wearable computing, Activity recognition, Sensory substitution