얼굴 생성 오토인코더를 이용한 단일 영상으로부터의 Valence 및 Arousal 추정

김도엽, 박민성, 장주용
광운대학교
dyubkim@kw.ac.kr, mspark@kw.ac.kr, jychang@kw.ac.kr

Estimation of Valence and Arousal from a single Image using Face Generating Autoencoder

Do Yeop Kim, Min Seong Park, Ju Yong Chang
Kwangwoon University

요 약

얼굴 영상으로부터 사람의 감정을 예측하는 연구는 최근 딥러닝의 발전과 함께 주목받고 있다. 본 연구에서 우리는 연속적인 변수를 사용하여 감정을 표현하는 dimensional model 에 기반하여 얼굴 영상으로부터 감정 상태를 나타내는 지표인 valance/arousal(V/A)을 예측하는 딥러닝 네트워크를 제안한다. 그러나 V/A 예측 모델의 학습에 사용되는 기존의 데이터셋들은 데이터 불균형(data imbalance) 문제를 가진다. 이를 해소하기 위해, 우리는 오토인코더 구조를 가지는 얼굴 영상 생성 네트워크를 학습하고, 이로부터 얻어지는 균일한 분포의 데이터로부터 V/A 예측 네트워크를 학습한다. 실험을 통해 우리는 제안하는 얼굴 생성 오토인코더가 in-the-wild 환경의 데이터셋으로부터 임의의 valence, arousal 에 대응하는 얼굴 영상을 성공적으로 생성함을 보인다. 그리고, 이를 통해 학습된 V/A 예측 네트워크가 기존의 under-sampling, over-sampling 방법들과 비교하여 더 높은 인식 성능을 달성함을 보인다. 마지막으로 기존의 방법들과 제안하는 V/A 예측 네트워크의 성능을 정량적으로 비교한다.

1. 서론

최근 딥러닝의 발전과 함께, 단일 얼굴 영상으로부터 사람의 감정을 인식하는 연구가 주목을 받고 있다. 사람의 감정 상태를 표현하기 위한 방법으로는 categorical model[1], dimensional model[2], facial action coding system(FACS)[3]을 들 수 있다. 간략히 말해서 categorical model 은 감정을 이산적인 값으로 표현하고, dimensional model 은 감정을 연속적인 값으로 표현하며, FACS 는 감정이 나타나는 얼굴의 미세한 근육 움직임을 표현한다. 본 연구에서 우리는 단일 얼굴 영상으로부터 dimensional model 기반의 valence/arousal(V/A)을 예측하는 방법을 제안한다. V/A 는 2 차원 공간에서 단위 원 안의 점으로 정의될 수 있으며, 이 경우 valence 와 arousal 은 각각 그 점의 x-좌표와 y-좌표에 대응된다. 여기서 valence 는 감정의 긍정적, 부정적 상태를 의미하고, arousal 은 감정의 세기를 의미한다.

대부분의 dimensional model 기반의 데이터셋들은 V/A 에 따른 데이터의 분포가 균일하지 않은 데이터 불균형(data imbalance) 문제를 가진다. 이는 이러한 데이터셋을 통해 학습된 V/A 예측 모델이 다양한 감정을 보이는 얼굴 영상에 대하여 감정을 올바르게 예측하지 못하게 만든다. 이러한 문제를 해소하기 위해 우리는 얼굴 영상 생성 네트워크를 제안한다.

제안하는 얼굴 생성 네트워크는 오토인코더(autoencoder) 구조를 가지며, 임의의 입력 V/A 에 대응하는 새로운 얼굴 영상을 생성한다. V/A 예측 네트워크는 이렇게 생성된 균일한 V/A 분포를 가지는 얼굴 영상을 사용하여 학습된다.

우리는 제안하는 얼굴 생성 네트워크를 정성적으로 평가한다.

그리고 우리는 제안하는 방법으로 학습된 V/A 예측 네트워크의 성능을 다양한 evaluation metric 을 사용하여 정량적으로 평가하며, 이를 통해 기존의 방법들보다 향상된 결과를 달성함을 보인다.

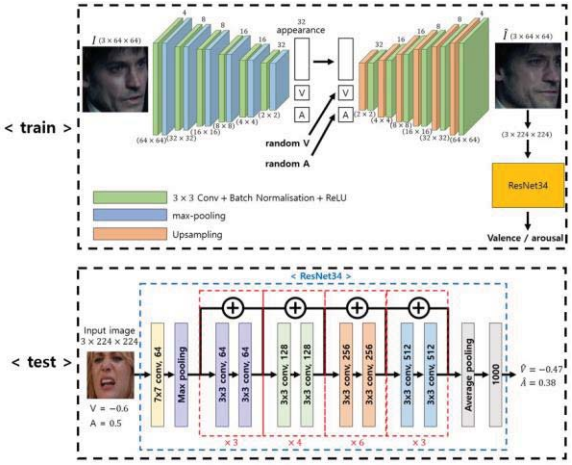

그림 1. 제안하는 방법의 개요

2. 제안하는 방법

우리는 제안하는 방법을 통해 기존 학습 데이터셋에 포함된 데이터 불균형을 해소하고자 한다. 이를 위해 우리는 얼굴 생성 오토인코더를 사용하여 V/A 예측 네트워크를 학습하는 방법을 제안한다. 얼굴 생성 오토인코더는 감정 분포가 불균일한 데이터셋으로부터 입력 얼굴 영상을 받고 uniform sampling 을 통해 얻어진 V/A 에 따라 변환된 얼굴 영상을 생성한다. 생성된 영상은 V/A 예측 네트워크의 학습을 위해 사용된다. 이로써 특정 감정에 집중된 데이터의 입력을 방지하고 V/A 예측의 성능을 개선할 수 있다. 그림 1 은 제안하는 방법의 개요를 보여준다.

1. 얼굴 생성 오토인코더

얼굴 생성 오토인코더는 그림 1 과 같이 인코더-디코더 (encoder-decoder) 구조를 가진다. 이는 $3 \times 64 \times 64$ 크기의 영상과 입력 V/A 로부터 동일한 해상도와 동일한 사람이지만 입력 V/A 에 맞게 변화된 표정의 새로운 영상을 생성한다. 인코더 E는 [3×3 conv, batch-normalization, ReLU, max-pooling] 모듈의 반복으로 구성된다. 디코더 D는 [upsampling, 3×3 conv, batch-normalization, ReLU] 모듈의 반복으로 구성된다. E는 입력 얼굴 영상으로부터 32 차원의 appearance feature z, valence v, arousal a를 포함하는 latent vector 를 출력한다. 네트워크가 이러한 latent vector 를 올바르게 출력하도록 학습하기 위해 우리는 학습 과정에서 서로 다른 표정을 가지는 동일한 사람의 영상을 사용한다. 그러한 두 영상이 I_i, I_j라고 가정하자. 먼저 우리는 E가 입력 얼굴 영상 I_i에 대한 V/A 를 올바르게 예측하도록 만들기 위해 다음과 같은 MSE loss 를 사용한다:

$$L_{\text{AE\_V}} = \|\hat{v}_i - v_i\|^2,\ L_{\text{AE\_A}} = \|\hat{a}_i - a_i\|^2, \quad (1)$$

여기서 $\hat{v}_i$, $\hat{a}_i$는 예측된 V/A 값을, v_i, a_i는 참값(ground-truth) V/A 값을 나타낸다. 다음으로 우리는 D가 E로부터 생성된 appearance feature z와 얼굴 영상 I_j에 대응하는 V/A 값인 v_j, a_j로부터 얼굴 영상 I_j를 올바르게 복원할 수 있도록 다음과 같은 reconstruction loss 를 사용한다:

$$L_{\text{Recon}} = \left\|\hat{I}_j - I_j\right\|^2, \quad (2)$$

여기서 $\hat{I}_j$ 는 D가 생성한 얼굴 영상을 나타낸다. 얼굴 생성 오토인코더를 학습하기 위한 전체 loss 함수는 다음과 같이 정의된다:

$$L_{\text{AE}} = \lambda_{\text{AE\_V}} L_{\text{AE\_V}} + \lambda_{AE\_A} L_{AE\_A} + L_{\text{Recon}}, \quad (3)$$

여기서 $\lambda_{\text{AE\_V}}$, $\lambda_{\text{AE\_A}}$ 는 각 loss 의 기여도를 조절하기 위한 가중치이다.

2. V/A 예측 네트워크

학습된 얼굴 생성 오토인코더는 균일한 데이터 분포를 가지는 데이터를 사용하여 V/A 예측 네트워크를 학습하기 위해 활용된다. 제안하는 V/A 예측 네트워크는 ImageNet 에 pre-train 된 ResNet34[4]에 기반한다. 분류(classification)를 위해 제안된 네트워크를 V/A 회귀(regression)에 사용하기 위해 우리는 ResNet34 의 마지막 softmax layer 를 제거하고 2 개의 node 를 가지는 linear layer 를 추가하였다. 네트워크의 구조는 그림 1 의 하단에 주어져 있다. 학습을 위해 샘플링 된 얼굴 영상은 얼굴 생성 오토인코더에 입력되며, E는 latent features z, v, a를 출력한다. 우리는 D의 입력으로 appearance z는 그대로 사용하며, V/A 는 다음과 같은 범위를 만족시키는 v_{new}, a_{new}를 uniformly sampling 하여 사용한다:

$$v_{new}^2 + a_{new}^2 \le 1. \quad (4)$$

D 에서 생성된 $3 \times 64 \times 64$ 의 영상은 $3 \times 224 \times 224$ 로 resize 되며, V/A 예측 네트워크는 resize 된 영상에 대응하는 V/A 예측 값 $\hat{v}_{\text{new}}$, $\hat{a}_{\text{new}}$를 출력한다. V/A 예측 네트워크를 학습하기 위한 loss 는 다음과 같이 정의된다:

$$L_{\text{V}} = \|v_{\text{new}} - \hat{v}_{\text{new}}\|^2,\ L_{\text{A}} = \|a_{\text{new}} - \hat{a}_{\text{new}}\|^2, \quad (5)$$

$$L_{\text{VA}} = L_V + L_A. \quad (6)$$

V/A 예측 네트워크는 L_{VA}를 사용하여 학습된다.

3. 실험 결과

우리는 제안된 V/A 예측 네트워크의 성능을 정량적으로 평가하기 위한 evaluation metric 으로 root mean square error (RMSE), correlation coefficient (CORR), concordance

correlation coefficient (CCC), sign agreement metric (SAGR)을 사용하였다. RMSE 는 작을수록 좋은 성능을 의미하고, CORR, CCC, SAGR 은 클수록 좋은 성능을 의미한다.

1. 데이터셋

본 연구에서 제안된 얼굴 생성 오토인코더와 V/A 예측 네트워크는 각각 AFEW-VA[5]와 AffectNet[6]으로 학습된다. AFEW-VA 는 다양한 영화 동영상으로부터 획득된 약 30,000 장의 영상으로 구성된다. 본 연구에서는 얼굴 생성 오토인코더를 학습하기 위해 24,757 장의 영상을 학습 데이터로 사용하고, 5,294 장의 영상을 테스트 데이터로 사용한다. AffectNet 은 인터넷에서 획득된 약 45 만장의 얼굴 영상으로 구성된 대규모 데이터셋이다. V/A 예측 네트워크를 학습하기 위하여 사람의 얼굴을 포함하지 않는 영상을 제외한 320,730 장의 영상을 학습 데이터로 사용하고, 4,500 장의 영상을 테스트 데이터로 사용한다. 네트워크에 입력할 얼굴 영상을 획득하기 위해서는 얼굴의 bounding box 가 필요하다. AffectNet 의 경우 데이터셋에서 제공하는 bounding box 를 사용하였고, 별도의 bounding box 를 제공하지 않는 AFEW-VA 의 경우 Dlib 의 딥러닝 기반 face detection model[7]을 사용하여 bounding box 를 획득하였다.

2. 구현 세부 사항

얼굴 생성 오토인코더를 학습하기 위해 Adam optimizer 를 사용하였고, learning rate, batch size, epoch 수는 각각 10^{-4}, 64, 50 으로 설정하였다. λ_V, λ_A는 모두 0.5 로 설정하였다. V/A 예측 네트워크의 학습을 위한 hyperparameter 는 얼굴 생성 오토인코더와 동일하게 설정하였다.

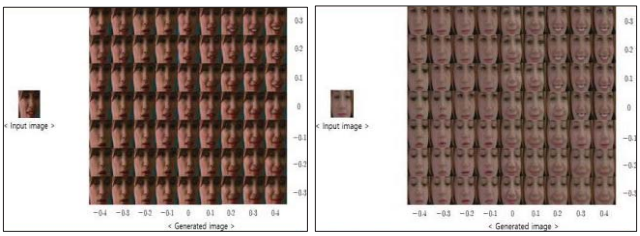

그림 2. 얼굴 생성 오토인코더가 입력 영상과 valence, arousal 값에 대응하는 새로운 영상을 생성한 결과

3. 임의의 V/A 에 대응하는 얼굴 영상 생성

그림 2 는 얼굴 생성 오토인코더가 입력된 얼굴 영상으로부터 새로운 영상을 생성한 결과를 보여준다. 입력 valence 와 arousal 은 각각 [-0.4, 0.4]와 [-0.3, 0.3]의 범위에서 0.1 의 간격으로 sampling 하였다. 우리는 V/A 값에 따른 감정의 변화를 잘 반영하는 얼굴 영상이 생성되었음을 확인할 수 있다.

표 1. 불균일한 데이터로 학습된 V/A 예측 모델과 제안하는 방법과의 정량적 성능 비교

	RMSE(V/A)	CORR(V/A)	CCC(V/A)	SAGR(V/A)
Original	0.406/ 0.359	0.612/ 0.533	0.573/ 0.501	0.742/ 0.752
Ours	**0.400/ 0.355**	**0.614/ 0.538**	**0.575/ 0.502**	**0.746/ 0.755**

4. 불균일한 데이터로 학습된 모델과의 비교

표 1 은 제안된 V/A 예측 네트워크를 불균일한 데이터셋으로 학습한 결과와 제안하는 방법으로 학습한 결과에 대한 정량적 비교를 보여준다. 모든 metric 에 대하여 제안된 방법이 불균일한 데이터셋으로 학습한 결과에 비하여 더 좋은 V/A 예측 성능을 보여준다. 이 실험으로부터 균일한 분포를 가지는 학습 데이터를 사용하는 것이 V/A 예측 성능에 도움을 준다는 것을 알 수 있다.

5. Data Sampling 방법에 따른 성능 비교

데이터 불균형 문제를 다루기 위해 흔히 사용되는 기존 방법으로는 over-sampling 과 under-sampling 을 들 수 있다. Over-sampling 은 특정 집단의 데이터 개수가 기준 threshold T_o 보다 작은 경우, T_o 가 될 때까지 같은 데이터를 복제한 후 사용하는 것을 의미한다. Under-sampling 의 경우, threshold T_u 보다 집단의 데이터 개수가 많으면 그 중에서 T_u 만큼만을 random sampling 하여 사용한다. 표 2, 3 은 각각 under-sampling 과 over-sampling 을 여러 개의 threshold 에 따라 적용하여 얻어진 데이터셋으로 V/A 예측 네트워크를 학습한 결과를 제안하는 방법의 결과와 함께 보여준다. 우리는 제안하는 방법이 기존의 sampling 방법을 사용한 결과에 비해 좋은 성능을 보임을 확인할 수 있다.

6. 기존 감정 인식 모델과의 성능 비교

표 4 는 제안하는 방법과 [6]에서 제안된 SVR, AlexNet 기반 모델과의 성능을 비교하여 보여준다. 3 가지 모델 중 가장 좋은 성능은 진하게 표시하였다. 제안하는 방법은 SVR 에 비하여 전체적으로 좋은 성능을 보인다. 하지만 AlexNet 과의 비교에서는 AlexNet 이 valence 예측에서 좋은 성능을 보이는 반면 제안하는 방법은 arousal 예측에서 더 나은 성능을 보인다는 것을 확인할 수 있다.

표 2. Under-sampling 방법과 제안하는 방법으로 학습한 V/A 예측 네트워크의 정량적 성능 비교

Methods		RMSE(V/A)	CORR(V/A)	CCC(V/A)	SAGR(V/A)
$T_u = 100$	Mean	0.415/0.368	0.596/0.510	0.506/0.484	0.719/0.731
	Std	0.031/0.034	0.035/0.032	0.034/0.029	0.031/0.035
$T_u = 500$	Mean	0.407/0.362	0.601/0.521	0.568/0.501	0.738/0.740
	Std	0.026/0.030	0.023/0.028	0.024/0.028	0.025/0.029
$T_u = 1000$	Mean	0.406/0.360	0.600/0.522	0.570/0.499	0.736/0.743
	Std	0.025/0.026	0.020/0.027	0.023/0.026	0.023/0.024
Ours	Mean	**0.400/0.355**	**0.614/0.538**	**0.575/0.502**	**0.746/0.755**
	Std	0.005/0.004	0.006/0.004	0.006/0.003	0.022/0.003

표 3. Over-sampling 방법과 제안하는 방법으로 학습한 V/A 예측 네트워크의 정량적 성능 비교

Methods		RMSE(V/A)	CORR(V/A)	CCC(V/A)	SAGR(V/A)
$T_o = 100$	Mean	0.408/0.364	0.601/0.519	0.561/0.485	0.738/0.731
	Std	0.020/0.016	0.021/0.021	0.021/0.018	0.023/0.020
$T_o = 500$	Mean	0.409/0.366	0.590/0.515	0.540/0.482	0.722/0.728
	Std	0.023/0.019	0.025/0.023	0.024/0.020	0.023/0.021
$T_o = 1000$	Mean	0.415/0.369	0.571/0.514	0.532/0.480	0.729/0.727
	Std	0.029/0.028	0.028/0.030	0.029/0.027	0.025/0.030
Ours	Mean	**0.400/0.355**	**0.614/0.538**	**0.575/0.502**	**0.746/0.755**
	Std	0.005/0.004	0.006/0.004	0.006/0.003	0.022/0.003

표 4. 제안하는 방법과 기존 모델과의 성능 비교

	AffectNet		Ours(V/A)
	SVR(V/A)	AlexNet(V/A)	
RMSE	0.550/0.420	**0.370**/0.410	0.400/**0.355**
CORR	0.350/0.310	**0.660/0.540**	0.615/0.538
CCC	0.300/0.180	**0.600**/0.340	0.575/**0.500**
SAGR	0.570/0.680	0.740/0.650	**0.745/0.755**

4. 결론

본 연구에서 우리는 단일 얼굴 영상으로부터의 감정 인식 문제에서 dimensional model 기반 데이터셋들이 가진 데이터 불균형을 해소하기 위해 임의의 V/A 값에 대응하는 새로운 얼굴 영상을 생성하는 얼굴 생성 오토인코더 및 이를 학습에 활용하는 V/A 예측 네트워크를 제안하였다. 실험 결과를 통해 우리는 제안된 얼굴 생성 오토인코더가 임의의 V/A 값에 대응하는 새로운 얼굴 영상을 성공적으로 생성함을 보였다. 또한 제안하는 방법으로 학습된 V/A 예측 네트워크는 기존의 over-/under-sampling 방법보다 정량적으로 높은 성능을 보였다. 마지막으로 기존의 감정 인식 모델과의 성능 비교를 통해 제안하는 방법이 정량적으로 우수한 V/A 예측 성능을 달성함을 확인하였다.

감사의 글

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2020 년도 문화기술연구개발 지원사업의 연구결과로 수행되었음. 본 연구는 2020 년도 과학기술정보통신부의 재원으로 정보통신기획평가원의 지원을 받아 수행되었음 (No. 2018-0-00735, UHD 방송 환경에서 콘텐츠에 대한 시청자의 반응 및 의도 기반 미디어 인터랙션 기술).

참고문헌

1. P. Ekman and W. V. Friesen, “Constants across cultures in the face and emotion,” Journal of Personality and Social Psychology, vol. 17, no. 2, 1971.
2. J. A. Russell, “A circumplex model of affect,” Journal of Personality and Social Psychology, vol. 39, no. 6, 1980.
3. P. Ekman and W. V. Friesen, “Facial action coding system,” 1977.
4. K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition." CVPR, 2016.
5. J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, “AFEW-VA database for valence and arousal estimation in-the-wild,” Image and Vision Computing, vol. 65, 2017.
6. A. Mollahosseini, B. Hasani, and M. H. Mahoor, “Affectnet: A database for facial expression, valence, and arousal computing in the wild,” IEEE Transaction on Affective Computing, vol. 10, no. 1, 2017.
7. D. King, "Max-margin object detection,” arXiv preprint arXiv:1502.00046, 2015.