💛

[12주차] Few-Shot Adversarial Learning of Realistic Neural Talking Head Models

0. Abstract

1️⃣ 대규모 데이터셋을 이용한 메타 학습

- 먼저 대규모 동영상 데이터셋을 이용해 **메타 학습**을 수행
- 이를 통해 새로운 사람의 **talking head 모델**을 빠르게 학습할 수 있는 능력을 갖추게 됨

2️⃣ 적대적 훈련을 통한 few-shot 학습

- 메타 학습 이후, 새로운 사람의 **few-shot 및 one-shot 학습을 적대적 훈련 문제로 수행**
- 이때 생성기와 판별기의 매개변수를 개인 맞춤으로 초기화하여, 수백만 개의 매개변수를 빠르게 조정할 수 있음

3️⃣ 사실적이고 개인화된 talking head 모델 생성

- 새로운 사람의 **매우 사실적이고 개인화된 talking head 모델**을 학습할 수 있음
- 초상화 그림에서도 이러한 talking head 모델 생성 가능

1. Introduction

- 본 논문은 개인화된 사실적인 talking head 모델을 생성하는 방법을 제안함
  - 이를 통해 **특정 개인의 사실적인 동영상 시퀀스를 합성**할 수 있음
  - 화상 회의, 멀티플레이어 게임, 특수 효과 산업 등에 실용적으로 활용 가능
- 그러나 사실적인 talking head 모델을 생성하는 데에 한계가 존재
  - 인간 얼굴의 높은 광도, 기하학적, 운동학적 복잡성: 얼굴뿐만 아니라 입 내부, 머리카락, 의복 등을 모델링해야 한다는 점
  - 인간 시각 체계의 높은 민감성: 인간 얼굴의 작은 오류에도 거부감을 느끼는 "**uncanny valley(불쾌한 골짜기)**"가 존재
  - 이미지 워핑 기술, GAN 등의 해결 방안을 제시했으나 역시 한계가 있었음

- 제안 방법: 본 논문에서는 소량의 사진만으로도 **사실적인 talking head 모델**을 생성할 수 있는 시스템을 제안함

  1️⃣ **메타 학습(meta-learning)**: 다양한 화자의 동영상 데이터셋으로 사전 학습하여 일반화된 모델을 만듦

  2️⃣ **적대적 학습(adversarial learning)**: 새로운 사람의 소량의 사진으로 빠르게 fine-tuning하여 사실적인 talking head 모델을 생성함

  ➡️ **단 한 장의 사진으로도 합리적인 결과를 얻을 수 있으며, 몇 장의 사진을 추가하면 개인화 수준이 향상됨. 또한 워핑 기반 시스템과 달리 다양한 포즈를 처리할 수 있음**

2. Related Work

얼굴 모델링

- 통계적 모델링 기법과 딥러닝을 이용한 **얼굴 모델링** 연구가 많이 진행되었음
- 그러나 talking head 모델링은 얼굴뿐만 아니라 머리카락, 목, 입 내부, 어깨 등 비 얼굴 부분도 모델링해야 하는 더 복잡한 문제

생성 모델링

- 본 논문의 시스템은 적대적 훈련(adversarial training)과 조건부 판별기(conditional discriminator) 등 **최근 생성 모델링 기법을 활용**함
- **메타 학습(meta-learning)** 단계에서 적응형 인스턴스 정규화(adaptive instance normalization) 기법을 사용

- 콘텐츠-스타일 분해(content-style decomposition) 아이디어를 활용, 포즈와 텍스처를 분리

메타 학습

- **모델 불변 메타 학습기(MAML)**는 메타 학습을 통해 이미지 분류기의 초기 상태를 얻고, 이를 이용해 소량의 데이터로 새로운 클래스를 빠르게 학습할 수 있음
- 본 논문에서도 이와 유사한 아이디어를 활용하지만, 구현 방식은 다름

적대적 메타 학습

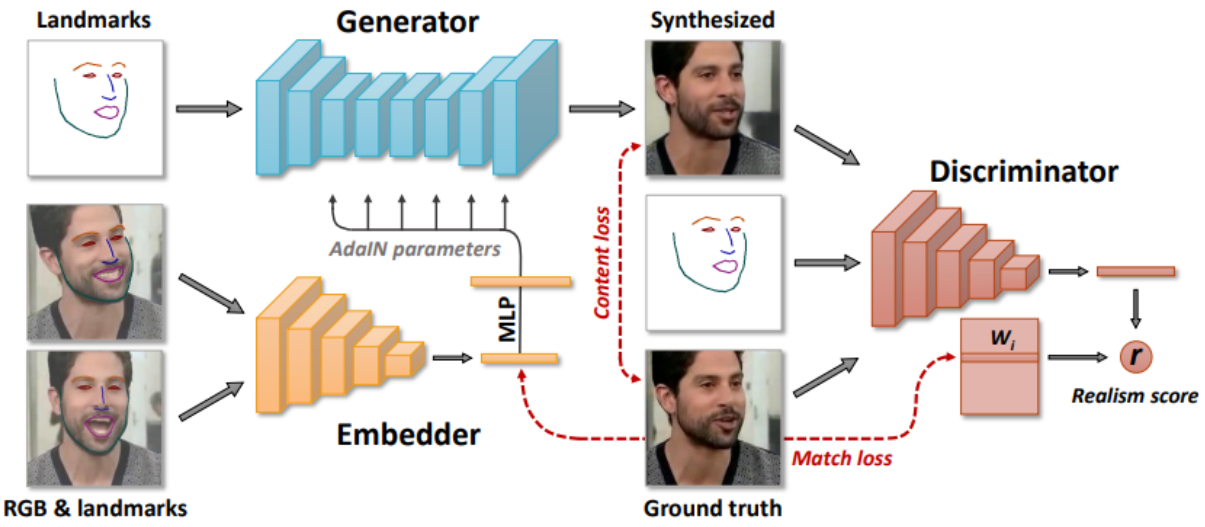
- 데이터 증강 GAN, MetaGAN, 적대적 메타 학습 등의 연구에서는 **메타 학습과 적대적 훈련을 결합하여 소량의 데이터로 새로운 클래스를 학습하는 방법**을 제안했음
- 이들 연구는 소량 샷 분류 성능 향상에 초점을 맞추었지만, 본 논문은 **이미지 생성 모델 학습에 적대적 미세 조정을 적용**함

텍스트-음성 합성

- 최근 텍스트-음성 합성 연구에서도 **소량 샷 학습**, **임베딩 네트워크**, **생성기 미세 조정** 등의 아이디어가 제안되었음
- 본 논문은 이를 talking head 모델링 문제에 적용하고, **적대적 학습과 메타 학습을 결합하는 방식**으로 구현함

3. Methods

3.1 Architecture and notation



- 입력 데이터
  - M개의 동영상 시퀀스 xi를 사용, 각 시퀀스에는 다른 사람의 talking head가 포함되어 있음
  - 각 프레임 xi(t)에 대해 얼굴 랜드마크 위치 정보 yi(t)를 사용하며, 이는 오프더셸프 얼굴 정렬 코드를 사용하여 얻음
  - 랜드마크 정보 yi(t)는 미리 정의된 색상으로 연결하여 3채널 이미지로 표현함
- 네트워크 구조

  1️⃣ **Embedder E(xi(s), yi(s); φ)**
  - 입력 프레임 xi(s)와 랜드마크 이미지 yi(s)를 N차원 벡터 êi(s)로 매핑
  - 메타 학습 과정에서 φ 파라미터를 학습하여, êi(s)가 포즈와 표정 변화에 불변하면서도 개인 정보를 포함하도록 함

  2️⃣ **Generator G(yi(t), êi; ψ, P)**
  - 랜드마크 이미지 yi(t)와 예측된 임베딩 êi를 입력으로 받아 합성 프레임 x̂i(t)를 출력
  - 일반적인 파라미터 ψ와 개인별 파라미터 ψî로 구성됨
  - 메타 학습 중에는 ψ만 직접 학습하고, ψî는 임베딩 êi와 학습 가능한 투영 행렬 P를 통해 예측함

  3️⃣ **Discriminator D(xi(t), yi(t), i; θ, W, w0, b)**
  - 입력 프레임 xi(t), 랜드마크 이미지 yi(t), 시퀀스 인덱스 i를 받아 하나의 현실성 점수 r을 출력
  - ConvNet 부분 V(xi(t), yi(t); θ)가 입력을 N차원 벡터로 매핑하고, 이를 이용해 W, w0, b 파라미터로 현실성 점수를 예측

3.2 Meta-learning stage

- 학습 단계

  1️⃣ K-shot 학습 에피소드를 시뮬레이션(실험에서 K=8)

  2️⃣ 무작위로 동영상 시퀀스 i와 해당 시퀀스의 프레임 t를 선택

  3️⃣ 추가로 K개의 프레임 s1, s2, ..., sK를 무작위로 선택

  4️⃣ 선택된 K개의 프레임에 대한 임베딩 êi(sk)를 평균하여 전체 시퀀스 i의 임베딩 êi를 계산함

  5️⃣ 계산된 임베딩 êi를 이용하여 프레임 t의 합성 이미지 x̂i(t)를 생성

- 손실 함수
  - **콘텐츠 손실 LCN**T: 실제 이미지 xi(t)와 합성 이미지 x̂i(t) 간의 perceptual similarity 손실
  - **적대적 손실 LADV**
    - 합성 이미지 x̂i(t)에 대한 판별기의 현실성 점수를 최대화
    - 특징 매칭 손실 LFM: 판별기의 특징 맵을 이용한 perceptual similarity 손실
  - **임베딩 매칭 손실 LMCH**: 임베더가 예측한 임베딩과 판별기의 임베딩 간 L1 차이를 최소화
- 판별기 학습
  - 판별기는 실제 이미지 xi(t)의 **현실성 점수를 높이고**, 합성 이미지 x̂i(t)의 **현실성 점수를 낮추도록** 학습됨
  - 이를 위해 hinge 손실 LDSC를 최소화하는 방향으로 판별기 파라미터를 업데이트

3.3 Few-shot learning by fine-tuning

- 입력 데이터
  - T개의 학습 이미지 x(1), x(2), ..., x(T), 해당 랜드마크 이미지 y(1), y(2), ..., y(T)
  - T는 메타 학습 단계의 K와 반드시 같을 필요 ❌
- 임베딩 계산
  - 메타 학습 단계에서 학습된 임베더 E(x, y; φ)를 사용하여 새로운 talking head 시퀀스의 임베딩 êNEW를 계산함
- 초기화
  - **생성기 G'(y(t); ψ, ψ')**는 메타 학습 단계의 생성기 G(y(t), êNEW; ψ, P)를 대체
  - ψ'는 새로운 개인 특화 파라미터로, êNEW와 P를 이용해 초기화됨
  - **판별기 D'(x(t), y(t); θ, w0, b)**의 ConvNet 부분 V(x(t), y(t); θ)와 편향 b는 메타 학습 단계의 결과로 초기화됨
  - w0는 메타 학습 단계의 Wi + w0와 êNEW의 합으로 초기화됨
- 손실 함수
  - **생성기**는 콘텐츠 손실 L'CNT(ψ, ψ')와 적대적 손실 L'ADV(ψ, ψ', θ, w0, b)를 최소화하도록 학습됨

    - **판별기**는 메타 학습 단계와 동일한 hinge 손실 L'DSC(ψ, ψ', θ, w0, b)를 최소화 하도록 학습됨
- 결과
    - 이와 같은 **fine-tuning 과정**을 통해 **새로운 사람의 talking head 시퀀스를 사실적으로 합성할 수 있음**
    - 메타 학습 단계의 초기화가 중요한데, 이를 통해 다양한 헤드 포즈와 표정에 대해 사실적인 이미지를 생성할 수 있음

3.4 Implementation details

- **Generator Network**
    - 제안된 생성기 네트워크 G(yi(t), êi; ψ, P)는 Johnson et al.의 image-to-image 변환 아키텍처를 기반으로 함
    - 다운샘플링과 업샘플링 레이어는 **Residual 블록**으로 대체되었으며, Batch Normalization은 **Instance Normalization**으로 대체되었음
    - 개인 특화 파라미터 ψî는 Instance Normalization 레이어의 **affine 계수**로 사용됨
- **Embedder** and **Discriminator**
    - 임베더 E(xi(s), yi(s); φ)와 판별기의 컨볼루션 부분 V(xi(t), yi(t); θ)는 유사한 네트워크 구조를 가짐
    - **Residual 다운샘플링 블록**으로 구성되며, 판별기에는 **추가적인 Residual 블록**이 포함됨
    - 출력은 **Global Sum Pooling과 ReLU를 거쳐 벡터화**됨
- 기타 구현 사항
    - 모든 레이어에 Spectral Normalization 적용
    - Self-Attention 블록이 일부 레이어에 추가됨
    - 콘텐츠 손실 LCNT는 VGG19와 VGGFace 네트워크의 중간 레이어 활성화 값을 사용
    - 특징 매칭 손실 LFM은 판별기 네트워크의 각 Residual 블록 출력을 사용
    - 임베딩 매칭 손실 LMCH 추가
    - 채널 수는 최소 64, 최대 512로 제한
    - 전체 네트워크 파라미터 수는 임베더 15M, 생성기 38M, 판별기 20M

    - 옵티마이저는 Adam이 사용되었으며, 학습률은 임베더/생성기 5e-5, 판별기 2e-4로 설정되었음

4. Experiments

- 데이터셋
    - VoxCeleb1 (256p, 1fps) 및 VoxCeleb2 (224p, 25fps) 데이터셋 사용
    - VoxCeleb2는 VoxCeleb1보다 약 10배 많은 동영상 포함함
- 평가 방법
    - 메타 학습 또는 사전 학습에 사용되지 않은 사람의 소량의 데이터(T개)로 fine-tuning
    - 동일한 시퀀스의 홀드아웃 부분으로 평가 수행
    - 50개의 VoxCeleb 테스트 동영상에서 32개의 홀드아웃 프레임 사용
- 평가 지표
    - Fréchet Inception Distance (FID): 지각적 사실성 측정
    - Structural Similarity (SSIM): 저수준 유사성 측정
    - Cosine Similarity (CSIM): 최신 얼굴 인식 모델의 임베딩 유사성 측정
- 사용자 평가
    - 동일한 사람의 실제 이미지 2개와 생성 이미지 1개로 구성된 트리플릿 제시
    - 사용자에게 가짜 이미지 찾기 요청
    - 사실성과 정체성 보존을 모두 평가할 수 있는 지표
- 비교 모델
    - X2Face: 워핑 기반 방법
    - Pix2pixHD: 직접 합성 방법
    - 제안 모델: 소량의 데이터로 fine-tuning하는 adversarial 학습 방식
- 비교 결과

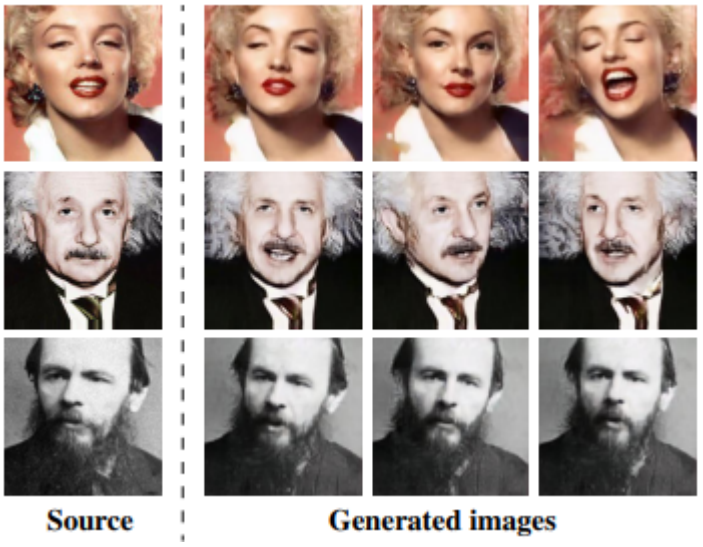
Method (T)	FID↓	SSIM↑	CSIM↑	USER↓
VoxCeleb1				
X2Face (1)	45.8	**0.68**	**0.16**	0.82
Pix2pixHD (1)	**42.7**	0.56	0.09	0.82
Ours (1)	43.0	0.67	0.15	**0.62**
X2Face (8)	51.5	**0.73**	**0.17**	0.83
Pix2pixHD (8)	**35.1**	0.64	0.12	0.79
Ours (8)	38.0	0.71	**0.17**	**0.62**
X2Face (32)	56.5	**0.75**	0.18	0.85
Pix2pixHD (32)	**24.0**	0.70	0.16	0.71
Ours (32)	29.5	0.74	**0.19**	**0.61**
VoxCeleb2				
Ours-FF (1)	**46.1**	0.61	**0.42**	**0.43**
Ours-FT (1)	48.5	**0.64**	0.35	0.46
Ours-FF (8)	42.2	0.64	**0.47**	0.40
Ours-FT (8)	**42.2**	**0.68**	0.42	**0.39**
Ours-FF (32)	40.4	0.65	**0.48**	0.38
Ours-FT (32)	**30.6**	**0.72**	0.45	**0.33**

- 1, 8, 32개의 프레임으로 fine-tuning하여 비교 실험 수행
- 각 테스트 프레임에 대해 동일한 사람의 다른 2개 프레임을 실제 이미지로 사용하여 트리플릿 구성

    - 정량적 평가 결과
        - X2Face와 Pix2pixHD가 제안 모델보다 SSIM, CSIM 지표에서 더 좋은 성능을 보임
        - X2Face는 L2 loss로 최적화되어 SSIM이 높고, Pix2pixHD는 인식 지표만 최적화되어 CSIM이 낮음
        - 그러나 이러한 지표들은 사용자 인지와 잘 부합하지 않음
        - 두 baseline 모델에서 "uncanny valley" 현상이 관찰되기 때문
    - 사용자 평가 결과
        - 4,800개의 트리플릿을 각 5명의 사용자에게 평가 요청
        - 제안 모델이 사실성과 개인화 측면에서 월등히 높은 성능을 보임
- 대규모 실험 결과

    - VoxCeleb2 데이터셋으로 두 가지 모델 학습
        - FF(feed-forward) 모델: 150 epoch, 임베딩 매칭 손실 없음
        - FT 모델: 75 epoch, 임베딩 매칭 손실 사용
    - FF 모델은 few-shot 학습 속도가 빠르고, FT 모델은 품질이 더 높음

    - FT 모델은 T=32 설정에서 사용자 평가 정확도 하한인 0.33을 달성하여 완벽한 성능 보임
- 퍼펫 조종 결과



    - 1-shot 설정으로 학습한 모델을 이용하여 사진 및 그림 퍼펫 조종
    - CSIM 기반으로 유사한 랜드마크 기하학을 가진 사람 선별
    - 다양한 퍼펫 조종 결과 제시

5. Conclusion

📍 **현재 한계점**

- 랜드마크 표현의 한계
    - 현재 사용된 랜드마크는 **시선 정보 포함** ❌
- 랜드마크 적응의 부재
    - 다른 사람의 랜드마크를 사용하면 **인물 불일치**가 발생
    - 이를 해결하려면 **랜드마크 적응**이 필요

📍 **향후 활용 방안**

- 자신의 대화형 인물 모델을 구동하는 용도에는 이미 높은 사실성을 제공함
- 다른 사람의 인물을 조종하는 **가짜 퍼펫 영상**을 만들려면 **랜드마크 적응**이 필요

논문에 대한 의견 및 의문점(꼭지)

➡️ 본 논문의 결론 부분에서, 한계점으로 랜드마크 적응 기술의 개선이 언급되었는데 이를 어떤 방향으로 개선하면 좋을지에 대해 논의해보고 싶음. 예를 들어 개인화된 랜드마크 모델링, 다양한 얼굴 특징 반영 등의 방법을 고려해볼 수 있을 듯 함