사전 학습된 GAN 으로부터의 전역 및 클래스별 의미론적 방향 추출 방법

김소라, 이민식
한양대학교

srk1995@hanyang.ac.kr, mleepaper@hanyang.ac.kr

Global and Class-Specific Semantic Directions via Pre-trained GANs

Sora Kim, Minsik Lee
Hanyang University

요 약

GANs 는 이미지 생성 분야의 혁신적인 발전을 가져왔지만, 아직 GAN 이 생성한 이미지 속에 함축되어 있는 의미론들을 명확하게 찾아내는 것은 어려운 문제이다. 기존 연구들을 통해 이 의미론들은 GAN 의 잠재공간상에 특정 방향의 형태로 나타난다는 것이 밝혀졌고, 잠재공간상의 한 샘플을 의미론 방향들로 움직이면 사람이 해석 가능한 형태로 이미지의 변화가 발생한다는 것이 밝혀졌다. 하지만 의미론 방향들을 찾는 방법은 아직까지 GAN 이 학습한 데이터셋의 모든 이미지 샘플에 걸쳐서 나타나는 공통적인 의미론을 찾고, 특정 클래스가 가지고 있을 의미론을 고려하지 않는다. 본 논문에서는 전역적 의미론과와 각 클래스별 의미론을 구분하여 찾고자 한다.

Ⅰ. 서 론

Generative adversarial networks (GANs) [1] 는 이미지 생성분야에서 많은 성공을 이뤘다. 특히 StyleGAN [2]은 AdaIN [3]을 통해 스타일 변화를 주며 고해상도의 이미지를 생성할 수 있다는 것을 보여주었다. 이를 토대로 다양한 연구에서는 StyleGAN 의 잠재공간을 분석하고자 하는 연구들이 많이 진행되었고, StyleGAN 의 잠재공간은 이전 연구인 PGGAN [4]에 비해 더 disentangle 하다는 것이 밝혀졌다.

한편, 최근 연구에서는 사전학습된 GAN 이 학습한 의미론을 추출하는 연구들이 진행되고 있다. 특히 [5]는 비지도학습 방법으로 서로 직교하는 의미론 방향을 찾았다. 하지만 이 방법은 사전학습된 GAN 의 학습데이터속 모든 이미지들이 가지고 있는 공통적인 의미론을 찾는 것에 집중하고 있고, 각 이미지가 속한 클래스가 가지고 있을 의미론을 고려하지 않는다. 따라서 생성된 이미지의 의미론을 정확하게 알아냈다고 할 수 없다. 따라서 본 논문에서는 사전 학습된 GAN 으로부터 데이터셋에 공통적으로 나타나는 전역적 의미론과 클래스별 의미론을 구분하여 추출하고자 한다.

Ⅱ. 본론

본논문에서는 잠재공간상에서의 의미론 변화를 만들어주는 변형자와, 원본 이미지와 변화된 이미지를 입력 받아 두 이미지의 차이로부터 변화량과 의미론의 방향을 예측하는 시프트 예측기를 통해 서로 직교하는 의미론의 방향을 찾아낸다. 학습 데이터의 클래스의 수가 N, 전체 방향의 수가 K 개, 변화의 범위가 ϵ 인 경우, 무작위로 1 부터 K 사이의 숫자 (k)를 추출하여 one-hot 벡터를 구성하고, ϵ 사이의 값 (ϵ)를 추출하여 이를 one-hot vector 에 곱하여 특정 방향으로 주어진 (ϵ)만큼 이동하는 역할을 하는 벡터 ($k\epsilon$) 를 구성한다.

변형자는 N+1 개의 학습 가능한 선형레이어 들로 구성되고, 각 클래스별 의미론 방향들과 전역적 의미론 방향을 학습한다. 클래스별 의미론 방향과 전역적 의미론 방향이 서로 같은 의미론을 추출하지 않도록 서로 직교하도록 구성하였다.

변형자를 통과해 나온 방향은 두 개인데, 특정 샘플 z 의 전역적 의미론 방향과 해당 샘플에 대응되는 클래스 방향이다. 특정 샘플 z 의 클래스는 사전학습된 분류기를 활용하여 획득했다. 이 두 방향을 샘플 z 에 더하여 잠재공간속에서 이동을 시킨 후 사전학습된 생성자로부터 변환된 이미지들을 얻어낸다.

시프트 예측기는 샘플 z 로부터 생성된 이미지와 변환된 이미지를 받아서 이 두 이미지의 차이를 통해 의미론 방향과 변화량을 찾는 모듈이다. 이 모듈은 변형자와 마찬가지로 전역적 의미론과 각 클래스의 의미론에 대응하는 방향과 변화량을 찾기 위해 N+1 개의 시프트 예측기를 정의한다. 각 시프트 예측기는 CNN 기반의 모델로 구성되어있다.

변형자와 시프트 예측기의 학습 식은 아래와 같다.

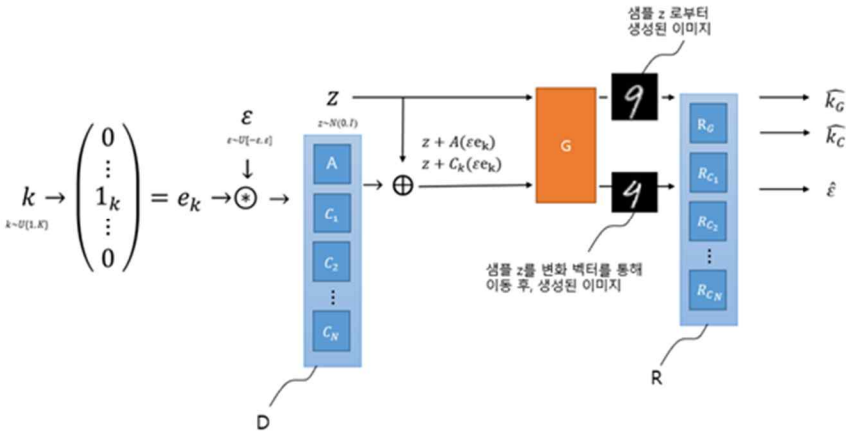

그림 1. 전역적, 클래스별 의미론 방향 및 변화량 추출 방법의 개요도.

$\min_{D_G,R_G}\left(L_{ds}\left(y_G-\hat{y}\right)+L_{reg}\left(\epsilon_G-\hat{\epsilon}_G\right)\right)+\min_{D_C,R_C}\left(L_{ds}\left(y_C-\hat{y}\right)+L_{reg}\left(\epsilon_C-\hat{\epsilon}_C\right)\right)$

이때, $y_G, y_C, \hat{y} \in \mathbb{R}^{K\times N+1}$ 이고, y_G 는 $[k\epsilon, 0, \ldots, 0]$ 이고, 첫번째 요소는 변형자의 입력으로 주어진 전역적 의미론의 변화의 방향과 변화량으로 구성된 벡터이다. y_C는 $[0, \ldots, k\epsilon, \ldots, 0]$으로 특정 클래스의 의미론의 변화의 방향과 변화량으로 구성된 벡터이다. 시프트 예측기를 통해 예측한 변화의 방향과 양 $\hat{y}$은 $[\hat{k}_G, \hat{k}_{C_1}, \ldots, \hat{k}_{C_N}]$이 된다. 그림 1 은 변형자와 시프트 예측기를 통해 예측되는 전역적 의미론의 방향과 클래스의 의미론 방향과 그에 대응하는 변화량을 보여준다.

그림 2 는 클래스 8 에 대한 클래스별 의미론 방향들에 대해 변화의 정도별로 변화하여 생성한 이미지들이다. 7 번째 이미지는 변화의 양이 0 인 경우 생성된 이미지에 해당하므로, 입력 이미지와 동일한 이미지를 만들어낸다. 왼쪽으로 갈수록 음의 변화를 나타내고, 오른쪽으로 갈수록 양의 변화를 나타낸다. 예를 들어, k=1 인 경우 숫자 8 의 두 원 부분이 서로 벌어지는 변화의 방향으로, 오른쪽으로 갈수록 변화의 정도가 커짐을 알 수 있다.

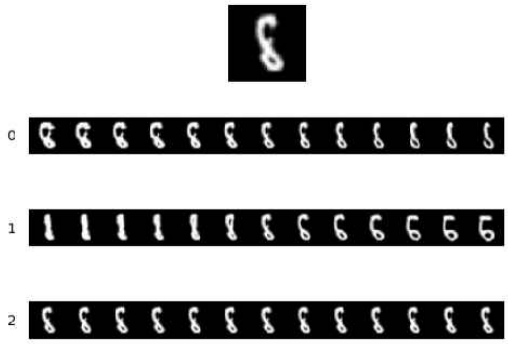

그림 2. 클래스 8 에 대한 의미론 방향들과 변화의 정도에 따라 생성된 이미지

Ⅲ. 결론

본 논문에서는 복수의 변형자와 시프트 예측기를 활용하여 전역적 의미론 방향과 클래스별 의미론 방향을 분리하여 추출하는 방법을 제안한다. 전역적 의미론 방향만을 찾는 기존의 방법과는 다르게 이미지 샘플의 클래스 정보를 고려한 방향을 찾음으로써 각 샘플이 가지고 있을 의미론 방향을 좀 더 자세하게 찾을 수 있다.

ACKNOWLEDGMENT

이 논문은 2024 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임. (No.RS-2022-00155885, 인공지능융합혁신 인재양성(한양대학교 ERICA))

참 고 문 헌

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in Neural Information Processing Systems, 27, 2014.

[2] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4401-4410, 2019.

[3] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, pages 1501-1510, 2017.

[4] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In Proceedings of International Conference on Learning Representations (ICLR) 2018. 2018.

[5] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In International conference on machine learning, pages 9786-9796. PMLR, 2020.