텍스트 기반 얼굴 이미지 생성 시스템

임동혁, 서용석
한국전자통신연구원(ETRI)

iammoni@etri.re.kr

Text-based Face Image Generation

Dong-Hyuck Im, Yongseok Seo
Electronics and Telecommunications Research Institute

요 약

본 논문은 텍스트 기반 이미지 생성 시스템을 제시한다. 먼저 VQGAN 모델을 사용하여 이미지를 표현할 수 있는 블록 기반의 양자화된 코드북을 학습하고, 학습된 코드북을 이용하여 이미지를 인코딩한 데이터와 텍스트 정보를 입력으로 GPT 모델을 학습한다. Multi-modal CelebA-HQ 텍스트/이미지 쌍 데이터셋을 이용하여 제안한 모델을 실험한 결과 텍스트를 입력으로 원하는 얼굴 이미지를 생성할 수 있었다.

Ⅰ. 서 론

DALL-E [1]와 같은 텍스트로 제시된 설명을 보고 이미지를 생성하는 텍스트 기반 이미지 생성 모델이 발표되고 있다. 본 논문에서는 텍스트 기반의 얼굴 이미지를 생성하기 위해 CNN 과 트랜스포머를 같이 사용하였다. CNN 을 이용하여 이미지를 벡터 양자화하고 잠재 공간으로 인코딩한다. 그리고, 트랜스포머를 이용하여 인코딩된 시퀀스에 대한 분포를 학습하고 새로운 이미지를 생성하였다.

Ⅱ. 본 론

본 논문에서는 이미지를 픽셀 단위가 아니라 블록 단위의 코드북 인덱스로 표현하고, 코드북 인덱스의 조합으로 텍스트 기반 이미지를 새롭게 생성하는 2 단계 접근 방식을 사용하였다.

첫 번째 단계 코드북 생성 과정에서는 VQGAN [2]을 이용하여 양자화된 코드북과 이미지를 코드화하는 인코더, 코드에서 이미지를 재구축하는 디코더를 한꺼번에 학습한다 [3]. VQGAN 모델 Q*는 다음의 방식으로 얻을 수 있다. FFHQ [4]와 CelebA-HQ [5] 데이터셋을 이용하여 학습하였다.

$$\mathcal{Q}^* = \arg\min_{E,G,\mathcal{Z}} \max_{D} \mathbb{E}_{x\sim p(x)} \Big[\mathcal{L}_{\mathrm{VQ}}(E,G,\mathcal{Z}) + \lambda \mathcal{L}_{\mathrm{GAN}}(\{E,G,\mathcal{Z}\},D) \Big]$$

L_{VQ} 는 코드북으로 이미지를 재구축할 때 원본과 차이가 적어지도록 설정된 손실 함수이고. L_{GAN} 은 생성한 이미지의 화질과 관련된 GAN 손실 함수이다. 이 두 가지 손실 함수의 합을 최소화하도록 학습을 진행한다.

코드북 학습이 완료되면 이미지를 양자화된 코드북 인덱스의 연속된 형태로 표현하고 텍스트 기반 이미지 생성 모델을 학습한다. 이미지를 코드북을 이용해서 인코딩하게 되면 텍스트 기반 이미지 생성 문제는 자연어 처리 분야의 다음 단어 예측 문제와 동일한 문제라 볼 수 있다. 우리는 최근 딥러닝 언어 모델 중 GPT-2 [6] 모델을 적용하였다. GPT 는 Generative Pre-trained Transformer 의 약자로, 트랜스포머 구조에서 디코더 부분을 활용하여 만든 모델이다. GPT 에서는 문장에서 주어진 단어들로부터 그 다음에 등장할 단어의 확률을 예측하는 방식으로 학습한다. 즉, 이전 인덱스 $x<i$ 가 주어졌었을 때 다음 인덱스 x_i 를 예측할 수 있도록 자기 지도 학습 방식으로 확률 분포 $p(x_i|x<i)$를 학습하게 된다.

$$\mathrm{p(x)} = \prod_{i=1}^{n} \mathrm{p}\left(x_i | x_i, \cdots, x_{i-1}; \theta\right)$$

학습에 사용되는 데이터는 Multi-Modal CelebA-HQ Dataset [7]를 활용하였다. Multi-Modal CelebA-HQ Dataset 은 공개된 데이터셋으로 30,000 장의 유명인 얼굴 이미지 데이터셋 CelebA-HQ [5]에 대한 텍스트 정보를 제공한다. 이미지들을 코드북으로 인코딩한 이미지 토큰과 텍스트 정보를 이용하여 학습을 진행하였다. 블록의 크기를 가로 16, 세로 16 픽셀로 하고 하나의 블록을 하나의 코드북 인덱스로 표현하면, 256x256 픽셀 이미지는 256 개의 연속된 이미지 토큰으로 표현할 수 있다.

```
<|startofimg|> [이미지를 설명하는 텍스트]
<|img|> [연속된 이미지 토큰]
<|endofimg|>
```

학습이 완료된 후, 얼굴 이미지를 생성하는 과정은 다음과 같다. 생성하고자 하는 이미지의 텍스트 정보를 <|startofimg|> [텍스트] <|img|>의 형태로 학습이 완료된 GPT 모델에 입력하면, 모델은 나머지 [연속된

이미지 토큰] <|endofimg|>를 예측하여 생성하게 된다. 얼굴 이미지를 생성하기 위해 이렇게 다음 토큰을 생성하면 또 그 다음 토큰을 생성하기 위해 방금 생성된 토큰을 다시 인풋으로 사용하는 자기회귀(auto-regressive) 방식으로 문장을 생성하게 된다. 생성된 이미지 토큰을 VQGAN 모델로 디코딩을 수행하면 입력한 텍스트에 해당하는 얼굴 이미지가 생성된다. 그림 1 은 텍스트 정보와 생성된 얼굴 이미지 쌍을 보여준다. 왼쪽에는 텍스트 정보, 오른쪽에는 텍스트 정보를 기반으로 생성한 얼굴 이미지이다.

그림 1 텍스트 정보(좌), 텍스트 기반 생성된 얼굴 이미지(우)

Ⅳ. 결론

본 논문에서는 VQGAN 과 GPT 모델을 함께 이용하여 텍스트 기반 얼굴 이미지 생성 모델을 학습하였다. 공개된 텍스트/이미지 쌍 데이터을 이용하여 제안한 방식을 적용한 결과, 텍스트 입력에 해당하는 고화질의 얼굴 이미지를 생성할 수 있음을 확인하였다.

ACKNOWLEDGMENT

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2022 년도 저작권기술 연구개발사업으로 수행되었음 (과제명: 교육 콘텐츠에 대한 인공지능 기반 저작권 침해 의심요소 검출 및 대체 재료 콘텐츠 추천 기술 개발, 과제번호: CR202104003)

참 고 문 헌

[1] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever. “Zero-Shot Text-to-Image Generation," arXiv:2102.12092v2. 2021.

[2] Patrick Esser, Robin Rombach, and Björn Ommer. “Taming Transformers for High-Resolution Image Synthesis." CVPR. 2021.

[3] 임동혁, 서용석. ”고화질 얼굴 이미지 생성 시스템,“ 한국통신학회 종합 학술 대회 (추계), 2021

[4] Tero Karras, Samuli Laine, Timo Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks,” arXiv:1812.04948. 2018

[5] Weihao Xia, Yujiu Yang, Jing-Hao Xue, Baoyuan Wu. “TediGAN: Text-Guided Diverse Face Image Generation and Manipulation," CVPR, 2021.

[6] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. ”Language Models are Unsupervised Multitask Learners," 2019.

[7] Tero Karras, Timo Aila, Samuli Laine, Jaakko Lehtinen. “Progressive Growing of GANs for Improved Quality, Stability, and Variation,” ICLR, 2018.