고위정보 특징을 이용한 향상된 노이즈 제거

이상호*, 유재영*, 곽노준
서울대학교
{shlee223, yoojy31, nojunk}@snu.ac.kr

요 약

우리는 영상과 음성에서의 노이즈 제거 문제(denoising)를 생성적 문제(generative task)로 풀고자 한다. 우리는 기존의 생성적 문제에서 state-of-the-art로 알려진 Generarive Adversarial Networks (GAN)의 성능을 더욱 향상시키기 위해, 영상 혹은 음성에 기본적으로 내재한 특징(low level feature) 이외에 영상의 사물에 대한 위치적 세분할 혹은 음성의 실제 언어적 의미 등 고위정보 특징(high leve feature)을 사전정보(prior)로 함께 사용하였다. 이를 통해 생성적 모델(generative model)이 좀 더 세밀한 정보를 쉽게 학습할 수 있을 것이라고 가정하였다. 우리는 기존의 state-of-the-art 신경망을 활용해 영상 혹은 음성에서 고위정보 특징을 추출해서 GAN과 함께 노이즈 제거를 수행하였고, 더욱 향상된 결과를 얻었다.

1. 서론

영상 혹은 음성은 그 샘플링(sampling) 과정이나 후처리(post processing) 과정에서 원치 않는 다양한 노이즈를 포함하여 생성될 수 있다. 그러나 최근 시각적, 청각적으로 더욱 쾌적한 영상, 음성에 대한 수요가 증가함에 따라 영상, 음성에서의 노이즈 제거(denoising) 혹은 아티팩트 제거(artifcat removal)의 중요성이 매우 부각되고 있다.

심층신경망(Deep neural network : DNN)은 보다 복잡한 비선형 시스템을 활용하여 실제 환경에서의 다양한 문제를 성공적으로 해결하고 있다. 따라서 노이즈 제거 문제에도 심층신경망을 적용하는 시도가 많아지고 있다. 기존의 DNN을 활용하는 방식은 일반적으로 회귀(regression)를 기반으로 Mean Squared Error(MSE) 비용 함수(cost function) 혹은 Mean Average Error(MAE) 비용 함수를 최적화하는 방법으로 학습을 수행한다.

그러나 위와 같은 MSE나 MAE 비용함수를 이용하는 방법은 잘 정의되지 않는 문제(ill-posed problem)에 대하여, 모델이 최적화 하려는 목표 데이터(target data)에 존재할 수 있는 가능한 다양한 가능성들의 평균을 찾아서 최적화하는 방향으로 학습되는 결과를 가져온다.

따라서 회귀 기반으로 학습된 모델은 총 오류값(error)은 적어지지만 데이터에 내제된 세밀함이 상쇄되어 사라질 수 있다. 이는 영상 혹은 음성의 생성문제(generative task) 에 있어 고주파수 성분이 제대로 생성되지 않거나 선명하지 않은 결과물을 생성하는 결과를 보일 수 있다.

최근의 생성적 심층신경망(Generative DNN) 방법론인 Generative Adversarial Networks(GAN)[2] 는 이런 문제를 해결하기 위한 좋은 방법으로 제시되고 있다. 영상이나 음성의 노이즈 제거 문제에 GAN을 적용할 시, 구분자 신경망(discriminator network)은 실제 데이터(real data)와 생성된 데이터(generated data)를 구분하는 방향으로 학습하며, 그 동안 생성자 신경망(generator network)은 이를 방해하는 방향으로 학습한다. 이 과정에서 생성자는 실제 샘플 분포가 존재하는 매니폴드(manifold)를 발견할 수 있을 것으로 기대한다.

이런 문제에 대하여 GAN 이외의 방법으로 구분적 모델(discriminative model)의 방법을 응용하는 사례도 있다. [3]의 저자는 이미지 자동 색칠(image colorization)문제에 있어, 이미지 분류(image classification)에 대하여 학습한 신경망의 특징맵을 생성적 신경망에 함께 사용함으로써 결과물의 품질을 더욱 상승시켰다. 예를 들어 이미지에 상단에 존재하는 평평한 표면을 학습데이터에 많이 존재했을 것으로 생각되어지는 하늘로 채색하기보다, 이미지 전체의 고위적 특징정보를 고려하여 실내 천장으로 채색하는 경우 등이 있다.

*표시의 저자들의 기여도는 같음

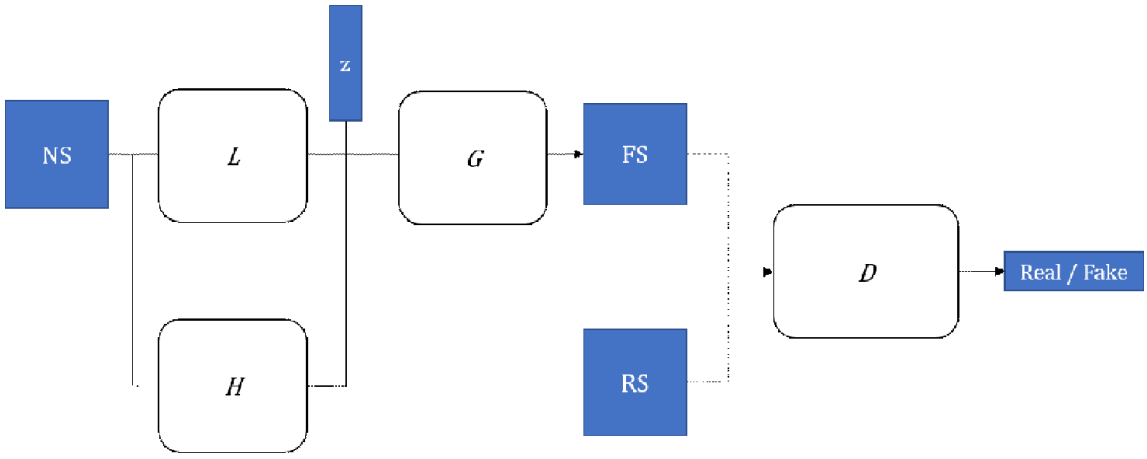


그림 1. 제안하는 네트워크 구조. 노이즈가 섞인 샘플(NS)는 신경망 L 로부터 기본정보 특징맵(low level feature), 신경망 H 로부터 고위정보 특징맵(high level feature map)이 추출된다. 이 특징맵들은 채널방향(channel wise)로 서로 결합하여, 생성자 신경망(G)에게 전달되고, G는 깨끗한 가짜 샘플(FS)을 생성한다. 이것은 깨끗한 실제 샘플(RS)과 쌍으로 구분자 신경망(D)를 학습시킨다.

우리는 이에 착안하여, 영상 혹은 음성의 노이즈 제거 문제를 GAN으로 접근하는 동시에, GAN의 생성자(generator)에 구분적 모델(discriminative model)의 특징맵을 함께 사용함으로써 이미지 노이즈 제거와 화자 음성 잡음제거의 성능을 향상시켜 보려 한다.

2. 제안하는 방법

우리는 GAN을 기반으로 한 노이즈 제거를 수행한다. 우리는 [6, 10]의 구조를 사용하여 생성자(generator), 구분자(discriminator) 구조를 설계했다. 여기서, 생성자 네트워크는 노이즈가 있는 입력 영상 혹은 음성뿐만 아니라 해당 입력에 해당하는 고위정보 특징(high-level feature)을 함께 이용한다. 우리는 여기에 추가할 고위정보 특징으로, 영상에 대해서는 영상 세분할 문제(image segmentation), 음성에 대해서는 음성-문자 변환 문제(speech-to-text translation)에 대하여 학습된 신경망의 결과 정보를 이용하기로 하였다. 즉, 노이즈가 있는 입력에 대하여 실제 그 입력에 해당하는 실측 자료(ground truth)를 예측하도록 지도 학습(supervised learning) 방식으로 신경망을 학습시키고, 이로부터 특징맵을 생성한다. 이 고위 정보 특징은 생성자 신경망이 노이즈가 있는 입력에서 노이즈가 제거된 결과물을 생성 할 때 일종의 사전정보로 작용할 수 있고, 결과적으로 그 세밀함을 향상시킬 수 있다.

그림 1 에서 볼 수 있듯이, 우리의 생성자 신경망은 기본정보 특징 신경망(L), 고위정보 특징 신경망(H), 생성자 신경망(G)로 구성된다. L은 입력에 대한 기본정보 특징맵(low level feature : LF)을 생성한다. H는 입력으로부터 이미지 세분할 또는 음성-문자 정보를 유추하도록 학습된 신경망으로, 고위정보 특징(high level feature : HF)을 생성한다. HF와 LF 그리고 무작위로 생성된 무작위 백터 z는 채널 축으로 연결되어서 G로 입력되며, G는 이를 이용해 결과물을 생성한다. 이 때 구분자(discriminator : D)는 G가 생성한 노이즈가 제거된 샘플(fake clean sample :FS)과 노이즈가 없는 실제 샘플(real clean sample: RS)을 가짜 혹은 진짜로 분류해낸다. L, G 그리고 D는 GAN (Generative Adversarial Net)[2]의 방법을 이용해 학습한다.

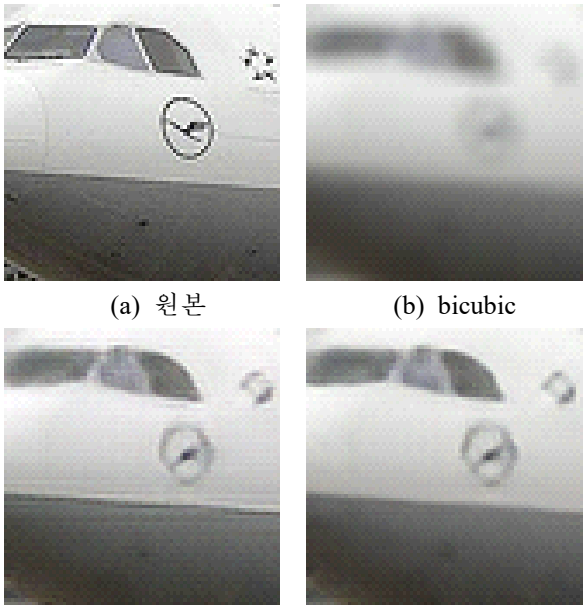
3. 실험 결과 및 분석

3.1 이미지 노이즈 제거 실험

이 실험에서 우리는 GAN을 이용한 이미지 초해상도(image super-resolution) 수행시 발생하는 노이즈를 제거할 수 있을 것으로 기대했다. 실험을 위한 이미지 세분할(image segmentation) 신경망은 FCN[9]의 fcn8s 구조로 pre-train 된 모델을 사용하였다. 이미지 초해상도를 수행할 GAN 신경망은 SRGAN [10]의 구조를 사용했다. 우리는 FCN를 이용하여 학습 데이터와 테스트 데이터에 해당하는 저해상도 이미지에서 각각 세분할 마스크(segmentation mask)를 추출하고, SRGAN의 출력에 채널 축으로 붙여서

각각 학습과 테스트를 수행하였다.

그림 2는 각각 원본 고해상도 이미지와 Bicubic interpolation, SRGAN 그리고 제안 알고리즘의 결과 생성물을 나타낸다. 제안 알고리즘은 bicubic 보다는 선명하고 SRGAN 보다는 선명하지 않지만 artifact가 더 적게 나타나는 것을 확인 할 수 있다.



(a) 원본 (b) bicubic

(c) SRGAN (d) Ours

그림 2. 다양한 방식(b-d)으로 수행한 이미지 초해상도 결과와 원본 이미지(a). 고위정보를 추가한 경우(d) GAN만을 통해 수행했을 경우보다(c) 노이즈가 적음.

최근 MSE를 기반으로 한 PSNR 수치가 최신수준의 이미지 초해상도 결과물에 대해 사람의 눈과 비슷한 판단기준이 될 수 없음을 보이는 연구들이 많이 등장하고 있다[10]. 따라서 최근 수준의 이미지 초해상도 결과물을 정량적으로 평가하는 방법은 현재 부재하다. 그러나 최근 생성된 이미지에 대한 특정 신경망의 특정 성능을 이미지 생성의 정량적 평가로 사용하는 시도들이 등장하고 있다.

표 1은 VOC2007 segmentation 데이터셋의 validation set에 대해서 원본 고해상도 이미지, Bicubic interpolation, SRGAN 그리고 이 연구의 제안 방법으로 노이즈를 제거한 생성물에 대하여 정량적 개선도를 실험한 결과이다. 이 수치는 생성한 이미지에서 FCN 신경망을 통한 이미지 세분할을 실제 수행하였을 때의 성능 수치이며, 이 수치가 실제 고해상도 이미지의 수치와 비슷할수록 이미지 초해상도가 더 잘 수행되었을 것으로 볼 수 있다.

표 1. FCN segmentation 신경망의 결과. 원본과 가까울수록 생성된 이미지의 품질이 좋다고 볼 수 있다.

	원본	Bicubic	SRGAN	Ours
FCN-Score	0.705	0.561	0.648	0.662

3.2 음성 잡음 제거 실험

실험을 위한 생성자 신경망과 구분자 신경망을 학습시키기 위해 NSDTSEA[7] 데이터셋을 사용하여 학습과 테스트를 수행하였다. NSDTSEA 데이터셋은 28명의 사람의 음성으로 이루어져 있으며, 깨끗한 음성과 노이즈를 입힌 음성 두가지로 나뉘어져 있다. NSDTSEA의 학습 데이터는 11000개 이상의 깨끗한 음성, 노이즈한 음성의 쌍으로 이루어져 있으며, 테스트 데이터로는 약 800개의 쌍으로 이루어져 있다. 학습의 연산량이 지나치게 늘어나는 것을 방지하기 위해서 학습데이터는 sliding window 방식으로 sub-sampling 되어 학습되었다. 테스트 단계에서는 전체 길이의 음성이 사용되었다.

실험을 위한 고위정보 특징맵 추출 신경망(H)는 [1]의 선행학습(pre-train) 된 모델을 사용하였다. 표 1은 NSDTSEA 데이터셋의 테스트셋인 824개의 오디오파일에 대해서 SEGAN[6]과 이 연구의 제안 방법으로 노이즈를 제거한 생성물에 대하여 정량적 개선도를 실험한 결과이다. 실험에 사용된 수치는 [6]에서 정량평가에 사용한 평가 척도를 참고하였으며, PESQ(Perceptual evaluation of speech quality), CSIG(Mean opinion score (MOS) prediction of the signal distortion attending only to the speech signal), CBAK(MOS prediction of the intrusiveness of background noise), COVL(MOS prediction of the overall effect), SSNR(Segmental SNR)을 사용하였다. 표 2에서 보이듯, 우리의 제안방법은 SEGAN보다 대체적으로 더욱 수치가 높았다.

표 2. NSDTSEA 데이터셋에 대한 노이즈한 원본과 SEGAN, 제안 방법에 대한 정량적 평가

	Noisy	SEGAN[6]	Ours
PESQ	1.97	1.99	2.06
CSIG	3.34	3.25	3.30
CBAK	2.44	2.78	2.83
COVL	2.63	2.59	2.66
SSNR	1.68	6.94	7.06

4. 결론

이 실험에서 우리는 GAN을 이용한 영상과 음성의 생성모델의 성능을 향상시키기 위해, 영상에 대해서는 위치 세분할 정보를, 음성에 대해서는 문자 언어정보를 고위정보 특징으로 함께 사용하여 실험하였다. 우리는 이 방법을 GAN을 이용한 이미지 초해상도와 마찬가지로 GAN을 이용한 음성 잡음 제거의 문제에 적용하여 기존의 방법이 더욱 개선됨을 보였다.

우리의 방법은 노이즈 제거 뿐만 아니라 다른 생성적 문제에 적용할 수 있으며, GAN 이외에 다른 생성모델에 쉽게 적용할 수 있다.

감사의 글

본 연구는 한국연구재단의 차세대 정보 컴퓨팅 기술 개발사업에 의해 진행되었음 (2017M3C4A7077582).

참고문헌

[1] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al. “Deep speech 2: End-to-end speech recognition in english and mandarin,” In International Conference on Machine Learning, pages 173–182, 2016.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. “Generative adversarial nets. In Advances in neural information processing systems,” pages 2672–2680, 2014.

[3] S. Iizuka, E. Simo-Serra, and H. Ishikawa. “Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification,” ACM Transactions on Graphics (TOG), 35(4):110, 2016.

[4] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. “Least squares generative adversarial networks,” arXiv preprint ArXiv:1611.04076, 2016.

[5] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. “Wavenet: A generative model for raw audio,” arXiv preprint arXiv:1609.03499, 2016.

[6] S. Pascual, A. Bonafonte, and J. Serrà. Segan: “Speech enhancement generative adversarial network,” arXiv preprint arXiv:1703.09452, 2017.

[7] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi. “Speech enhancement for a noiserobust text-to-speech synthesis system using deep recurrent neural networks,” In INTERSPEECH, pages 352–356, 2016.

[8] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A. "The PASCAL Visual Object Classes (VOC) Challenge", International Journal of Computer Vision, 88(2), 303-338, 2010

[9] J. Long, E. Shelhamer, and T. Darrell. "Fully convolutional networks for semantic segmentation", CoRR, abs/1411.4038, 2014.

[10] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. P. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. "Photo-realistic single image super-resolution using a generative adversarial network. CoRR, abs/1609.04802, 2016.